A Survey of Day-Night Illumination Domain Translation for Outdoor Vision Covering 30 Methods 22 Datasets and Evaluation Protocols

Md Shadab Alam[1*], Priyanshu Singh[2] and Pavlo Bazilinskyy[1]

[1*]Eindhoven University of Technology, Eindhoven, 5612DS, The Netherlands.
[2]Dr. B.C Roy Engineering College, Durgapur, 713206, India.

*Corresponding author(s). E-mail(s): m.s.alam@tue.nl;
Contributing authors: priyanshu.asn2003@gmail.com; p.bazilinskyy@tue.nl;

Abstract

Day-night appearance shift degrades vision for driving and surveillance. Low illumination, mixed lighting, glare, and sensor noise weaken cues for detection, segmentation, localisation, and tracking. We survey illumination domain translation for images and video, focusing on day to night and night to day mapping that changes illumination while preserving geometry, semantics, and temporal coherence. We relate illumination modelling and colour transfer to learning based methods, and introduce a constraint centric taxonomy linking supervision, five domain gap factors, and five families of constraints and priors to typical failure modes. Using this taxonomy, we organise 30 representative methods and summarise 22 datasets. We also report an artefact availability audit of 34 published methods: 29 release code, 22 provide pretrained weights, 21 specify licences, and 19 provide reproducibility packages. Finally, we recommend evaluation spanning perceptual quality, semantic preservation, downstream utility, and temporal stability.

Keywords: Image-to-Image Translation, Domain Adaptation, Semantic Preservation, Outdoor Vision, Reproducibility

1 Introduction

Computer vision underpins a wide range of deployed systems, including automated driving, surveillance, and large scale recognition pipelines (Du, Shi, Zeng, Zhang, & Mei, 2022; L. Liu et al., 2020; Ma, Ouyang, Simonelli, & Ricci, 2024). However, in practice, performance degrades sharply under adverse imaging conditions. At night, poor illumination, short exposure, low contrast, and sensor noise suppress or distort the visual evidence that modern perception models rely on, including object boundaries, textures, road markings, and small or distant targets, leading to substantial accuracy drops during real world operation (C. Li et al., 2021; J. Liu, Xu, Yang, Fan, & Huang, 2021). A central difficulty is the domain gap between the conditions represented in standard training and evaluation data and those encountered in low illumination deployment, where appearance statistics, signal to noise characteristics, and visibility of semantic cues differ markedly from daytime imagery. This gap persists and is often amplified because widely used corpora and benchmarks are dominated by well lit images, whereas dedicated nighttime datasets remain comparatively small and difficult to collect, and their dense annotations are considerably more expensive, which limits coverage and encourages models to overfit to daytime appearance (Cordts et al., 2016; J. Liu et al., 2021; Neumann et al., 2018; Yu et al., 2020).

An approach to mitigating this gap is Illumination Domain Translation (IDT). We review recent methods for outdoor imagery and discuss how constraint design and evaluation practice affect practical utility. The scope and terminology used in this survey are defined in subsection 3.1. Previous work shows that IDT can attenuate adverse effects such as low brightness, reduced contrast, and sensor noise by producing translated imagery that better aligns with the conditions seen during training (Pang, Lin, Qin, & Chen, 2021). In the specific context of night to day (N2D) and day to night (D2N) translation, paired and unpaired learning frameworks (Isola, Zhu, Zhou, & Efros, 2017; Zhu, Park, Isola, & Efros, 2017) illustrate how cross domain mappings can reduce the mismatch between training data and deployment conditions. From a data perspective, the difficulty of capturing and annotating low light images at scale (J. Liu et al., 2021) motivates the widespread use of weakly supervised and unpaired approaches, as well as outdoor specific translation systems (E. Lee & Kang, 2021), which avoid the need for perfectly aligned day and night image pairs.

Despite these advantages, N2D translation remains challenging, as the mapping from night time to daytime appearance is not uniquely determined by the input. Low illumination and short exposure amplify sensor noise and reduce contrast, erasing boundaries and small objects that are critical to outdoor perception (J. Liu et al., 2021). Localised light sources introduce strong spatial illumination discontinuities (Jobson, Rahman, & Woodell, 1997), while mixed artificial spectra violate standard colour constancy assumptions (Gehler, Rother, Blake, Minka, & Sharp, 2008). As a result, multiple daytime interpretations may be compatible with a single night time observation, necessitating strong priors, as already recognised in classical illumination and reflectance formulations (Land & McCann, 1971). From a probabilistic perspective, this ambiguity manifests itself as perceptual uncertainty and multimodality (Blau & Michaeli, 2018; T. Wang et al., 2022), explaining why translated outputs may appear visually plausible but remain semantically incorrect. These issues motivate our problem formulation, since perceptual improvements do not reliably translate into downstream performance gains (R. Zhang, Isola, Efros, Shechtman, & Wang, 2018), and in the video setting explicit temporal coherence constraints are required to prevent flicker and inconsistent structure between frames (T.-C. Wang, Liu, Zhu, Liu, et al., 2018).

Building on the trade off between perceptual quality and fidelity highlighted by Blau and Michaeli (2018), and related observations by R. Zhang et al. (2018), this review addresses a key gap in the N2D translation literature: while many methods report improved visual appearance, it is often unclear which design choices reliably preserve scene content and structure under extreme changes in illumination, and how such choices should be evaluated for practical use. Accordingly, we organise prior work around constraints and priors intended to preserve semantics, geometry, and, for video, temporal coherence, rather than optimising visual realism alone. To motivate this organisation, we relate classic principles, including illumination and reflectance separation (Land & McCann, 1971) and colour transfer formulations (Reinhard, Adhikhmin, Gooch, & Shirley, 2002), to modern generative objectives that operationalise these ideas as explicit constraints.

We cover physics motivated and decomposition based formulations, such as illumination and atmospheric simulation (Lengyel, Garg, Milford, & van Gemert, 2021) and disentanglement orientated translation (Lan, Zhao, & Li, 2023), as well as supervised methods (Lakmal, Dissanayake, & Aramvith, 2024), unpaired methods (E. Lee & Kang, 2021), semantic aware objectives for structure preservation (Schutera, Hussein, Abhau, Mikut, & Reischl, 2020), multimodal and fusion orientated strategies (Yang, Sun, Lou, Yang, & Zhou, 2023), and temporal formulations for video (T.-C. Wang, Liu, Zhu, Liu, et al., 2018). Guided by the evaluation concerns raised by Blau and Michaeli (2018) and R. Zhang et al. (2018), we also systematise datasets and protocols that assess perceptual quality alongside preservation and task utility, and we complement the methodological review with an artefact availability audit spanning code, model weights, training configurations, evaluation scripts, and licence terms.

Unlike previous reviews that focus primarily on low light enhancement (Guo, Ma, García-Fernández, Zhang, & Liang, 2023), image colourisation (Anwar et al., 2025; S. Huang, Jin, Jiang, & Liu, 2022), or generic image to image translation (Hoyez, Schockaert, Rambach, Mirbach, & Stricker, 2022; Saxena & Teli, 2021), this article treats IDT as a distinct problem for outdoor vision (Anoosheh, Sattler, Timofte, Pollefeys, & Van Gool, 2019; Dai & Gool, 2018; Sakaridis, Dai, & Gool, 2019). In this setting, translation errors can alter semantics in ways that are not captured by perceptual metrics alone (Cherian & Sullivan, 2019), mapping is often multimodal (X. Huang, Liu, Belongie, & Kautz, 2018), and video introduces temporal instability that can undermine downstream pipelines (T.-C. Wang, Liu, Zhu, Liu, et al., 2018). We therefore analyse failure modes that can be missed by perceptual metrics and organise methods according to the constraints used to control these failures. Finally, we foreground reproducibility and deployability through a structured artefact availability audit, a focus that is often underemphasised in previous reviews (Guo et al., 2023; Pitié, 2020) and aligns with recent calls for greater openness and transparency in automotive user research (Ebel et al., 2024).

To differentiate this survey from previous work, we make four contributions. First, we introduce a constraint centric taxonomy that links supervision, sources of domain gap, families of constraints and priors, and recurring failure modes. Second, we provide structured coverage of 30 methods and 22 datasets, and we state explicit inclusion criteria to make the scope clear and reproducible. Third, we propose a four part evaluation protocol based on evidence of perceptual quality, structural fidelity, downstream task impact, and temporal stability, abbreviated as P, S, D, and T, and we highlight common reporting pitfalls, for example using PSNR or SSIM without ensuring geometric alignment. Fourth, we report an artefact availability audit that records what was checked, the audit snapshot date, and a precise definition of availability.

1.1 Problem formulation and implications

Let $\mathcal{X}_n$ and $\mathcal{X}_d$ denote the distributions of RGB images at night and daytime, respectively. The N2D and D2N translations seek a mapping

$$G : \mathcal{X}_n \to \mathcal{X}_d \tag{1}$$

(or its inverse) such that the translated output matches the target domain while remaining faithful to the input scene.

Organising lens (scope control). We use the following decomposition as a unifying *view* of the literature: many learning-based methods can be interpreted as combining a domain-alignment term with a subset of constraint/prior terms that mitigate specific failure modes. This is not a claim that all approaches optimise an identical objective, nor that any objective guarantees semantic faithfulness under severe ambiguity.

Most learning-based approaches can be interpreted as combining a *domain-alignment* term which encourages outputs to match the target illumination domain with additional *constraint/prior* terms that mitigate specific failure modes (e.g. semantic drift, hallucinated structure, and, for video, temporal instability):

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} \\ & + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{phy}}\mathcal{L}_{\text{phy}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}\end{aligned} \tag{2}$$

Here $\mathcal{L}_{\text{adv}}$ provides distributional alignment, typically via an adversarial objective (Isola et al., 2017); $\mathcal{L}_{\text{cyc}}$ enforces the consistency of the cycle or reconstruction (Zhu, Park, et al., 2017); $\mathcal{L}_{\text{id}}$ promotes the preservation of identity (Zhu, Park, et al., 2017); $\mathcal{L}_{\text{sem}}$ encodes semantic or task-consistency constraints through labels or task networks (Hoffman et al., 2018); $\mathcal{L}_{\text{phy}}$ captures physics-informed priors (e.g. illumination and image-formation structure) (Land & McCann, 1971; Lengyel et al., 2021); and $\mathcal{L}_{\text{temp}}$ enforces temporal coherence for video translation (T.-C. Wang, Liu, Zhu, Liu, et al., 2018). Nonnegative coefficients $\lambda_{\text{adv}}, \lambda_{\text{cyc}}, \lambda_{\text{id}}, \lambda_{\text{sem}}, \lambda_{\text{phy}}, \lambda_{\text{temp}}$ are method-specific weights that place a relative emphasis on alignment versus constraint terms (Isola et al., 2017; T.-C. Wang, Liu, Zhu, Liu, et al., 2018; Zhu, Park, et al., 2017). As an organising lens, this formulation provides a compact index into the method space: individual approaches can be understood as instantiating different subsets (and variants) of these terms, leading to different trade-offs between visual plausibility, structural preservation, semantic faithfulness, and temporal stability.

1.1.1 Alignment term in paired and unpaired settings

The alignment component depends on the available supervision. When paired (or pseudo-paired) samples (x_n, x_d) are available, alignment is typically enforced by direct reconstruction (optionally complemented by perceptual terms):

$$\mathcal{L}_{\text{align}}^{\text{paired}} = \mathbb{E}_{(x_n,x_d)}\Big[\ell_{\text{rec}}\big(G(x_n), x_d\big) + \alpha\,\ell_{\text{perc}}\big(G(x_n), x_d\big)\Big] \tag{3}$$

where, ℓ_{rec} is commonly an ℓ_1 or ℓ_2 loss and ℓ_{perc} is a learnt perceptual distance. Adversarial terms can be added to improve realism in the target domain (Isola et al., 2017).

When supervision is unpaired, alignment is typically distributional rather than pixel-wise:

$$\mathcal{L}_{\text{align}}^{\text{unpaired}} = \mathcal{L}_{\text{adv}}\big(G; \mathcal{X}_n \to \mathcal{X}_d\big) \Big(+\ \mathcal{L}_{\text{adv}}\big(F; \mathcal{X}_d \to \mathcal{X}_n\big)\Big) \tag{4}$$

where, $\mathcal{L}_{\text{adv}}$ denotes the adversarial alignment between translated outputs and the target-domain distribution (Isola et al., 2017). In this regime, additional constraints such as cycle/identity, semantic/task, physics-based, and temporal terms provide the primary mechanism for preserving scene structure under severe illumination shift.

A defining characteristic of IDT is that the mapping is often underdetermined in low light, where noise, saturation, and occlusions remove or corrupt the evidence of the scene (J. Liu et al., 2021). Consequently, for a night observation $x_n \in \mathcal{X}_n$, the conditional distribution $p(x_d \mid x_n)$ is frequently multimodal. Importantly, this implies there may be multiple daytime (or nighttime) renderings that are consistent with the same observation, so "correctness" is better viewed as satisfying faithfulness constraints than matching a unique ground-truth output. Deterministic models therefore typically return a single plausible solution within this space (X. Huang et al., 2018; H.-Y. Lee, Tseng, Huang, Singh, & Yang, 2018; Zhu, Zhang, et al., 2017). This can yield outputs that appear visually convincing while deviating from the underlying scene semantics, consistent with the perception distortion trade-off (Blau & Michaeli, 2018). The risk is amplified by assumptions that are routinely violated at night, including smooth illumination in the presence of localised light sources (Jobson et al., 1997), reduced signal-to-noise ratios under short exposure (C. Li et al., 2021), limited visibility of small or distant semantic cues (Neumann et al., 2018), and Lambertian reflectance and colour constancy under mixed artificial spectra (Gehler et al., 2008). In safety-critical settings, these considerations favour conservative transformations that preserve structure and object integrity over aggressive appearance normalisation.

These observations motivate the central organising principle of this survey. Progress in N2D and D2N translation is best understood through the constraints and priors used to control semantic drift, hallucinated structure, and temporal instability, and through evaluation protocols that test these failure modes directly.

2 Foundations before deep generative models

Before deep learning based image translation, approaches that addressed day and night appearance variation were dominated by classical illumination modelling and colour processing, including Retinex style lightness and colour constancy formulations, multiscale illumination correction, and statistical colour transfer (Jobson et al., 1997; Land & McCann, 1971; Reinhard et al., 2002). The common objective was to reduce sensitivity to illumination by correcting low brightness, stabilising colour, or normalising lighting so that the scene more closely resembles a well lit observation, with related extensions to video through time dependent illumination normalisation (Matsushita, Nishino, Ikeuchi, & Sakauchi, 2004; Toet

& Hogervorst, 2012). Retinex inspired illumination correction and intrinsic image style decompositions are central in this line of work, and they provide useful context for understanding how modern translation methods impose constraints to preserve scene content under extreme illumination change.

2.1 Illumination and colour normalisation

Early work on coping with day and night appearance variation focused on normalising image formation effects rather than learning cross domain mappings. A common first line of attack was the manipulation of contrast and dynamic range, including local histogram based methods such as contrast limited adaptive histogram equalisation and related local contrast operators (Zuiderveld, 1994). Closely related work in tone mapping for high dynamic range imagery decomposed images into large scale illumination and fine detail components using edge preserving filtering, then compressed the illumination component while retaining detail (Durand & Dorsey, 2002). Although these operators do not explicitly model D2N translation, they shaped many later enhancement and normalisation baselines by formalising practical ways to boost visibility while limiting halos and contrast artefacts.

A second classical line addressed colour constancy by estimating the scene illuminant and correcting chromatic shifts so that object colours remain stable across lighting changes. The Grey World hypothesis introduced by Buchsbaum (1980) estimates illumination from global statistics, while Forsyth (1990) formulated a principled approach to constraining illuminant estimates using feasible colour gamuts. Later variants improved robustness by relaxing the strict Grey World assumption through intermediate norms (Finlayson & Trezzi, 2004) and by using derivative statistics rather than raw intensities, as in the Grey Edge family (Van De Weijer, Gevers, & Gijsenij, 2007). These methods provide interpretable mechanisms for separating lighting effects from scene appearance, and they motivated distribution alignment style reasoning that later appeared in learning based translation objectives.

A third family used reflectance illumination decomposition ideas to construct illumination stable representations. Intrinsic image formulations explicitly describe an observed image as reflectance modulated by illumination and seek to recover components that are more invariant to changes in light (Barrow, Tenenbaum, Hanson, & Riseman, 1978). In recognition orientated settings, quotient based normalisation was explored as a practical approximation to removing lighting effects, for example through the Self Quotient Image representation for illumination robust matching (H. Wang, Li, Wang, & Zhang, 2004). For video, illumination can also be treated as a time varying signal; Matsushita et al. model illumination in a low dimensional subspace to compensate lighting fluctuations and cast shadows in fixed camera sequences (Matsushita et al., 2004).

Retinex theory integrates these perspectives by computing lightness from relative luminance ratios rather than absolute intensities, supporting colour constancy and reduced dependence on the illumination field (Land & McCann, 1971). The practical Retinex variants then introduced explicit engineering choices to improve stability in real imagery. Multiscale Retinex with Colour Restoration applies centre surround operations at multiple spatial scales and adds a colour restoration stage to counteract desaturation, delivering improved dynamic range compression while retaining chromatic information (Jobson et al., 1997). Subsequent work reformulated Retinex with explicit optimisation, for example through a variational framework and reduced complexity solvers (Elad, Kimmel, Shaked, & Keshet, 2003; Kimmel, Elad, Shaked, Keshet, & Sobel, 2003), and edge preserving filtering was used to improve handling of illumination boundaries and suppress noise in dark regions (Elad, 2005). More recent fast Retinex style baselines estimate illumination using simple statistics in HSV space and recover an approximate reflectance at low computational cost, for example RBFA (S. Liu, Long, He, Li, & Ding, 2021).

Taken together, these classical strands establish the key ingredients that modern day and night translation systems must control: visibility and dynamic range, chromatic stability under mixed illumination, and separation of scene structure from illumination effects, with additional stability requirements when processing video.

2.2 Limitations and connection to modern translation

Classical illumination and colour normalisation methods provide valuable physical insight, but they typically rely on assumptions that are frequently violated in real outdoor night scenes. These assumptions include spatially smooth illumination, limited dynamic range, and stable colour constancy (Buchsbaum, 1980; Durand & Dorsey, 2002; Forsyth, 1990; Jobson et al., 1997; Land & McCann, 1971). Localised light sources, saturated highlights, mixed lighting spectra, and sensor noise can therefore lead to unstable corrections and limited recovery of semantically meaningful detail in severely exposed regions (S. Liu et al., 2021; Neumann et al., 2018; Sakaridis et al., 2019). In addition, classical pipelines are largely agnostic to semantic structure, which limits their suitability for safety critical applications where object identity and layout must be preserved (Bhattacharjee, Kim, Vizier, & Salzmann, 2020; Schutera et al., 2020). For video, classical normalisation can reduce global illumination variation, but it generally does not enforce stable content translation between frames in complex motion and mixed lighting (Matsushita et al., 2004).

A useful link to modern methods is provided by a simplified view of image formation.

$$I(x) = R(x) \cdot L(x) \tag{5}$$

where, $I(x)$ is the observed image, $R(x)$ denotes reflectance that encodes the structure of the illumination invariant scene, and $L(x)$ denotes the illumination field. Retinex and intrinsic image formulations seek to estimate or normalise $R(x)$ by imposing hand crafted priors on $L(x)$, but intrinsic decomposition is severely under constrained without strong assumptions (Barron & Malik, 2012; Barrow et al., 1978). Learning based translation can be interpreted as replacing such hand crafted priors with data driven constraints that regularise the mapping while preserving content.

This continuity is visible in multiple design patterns. Retinex motivated separation of illumination and reflectance reappears in learning based methods that disentangle content from illumination or style through explicit decomposition and representation constraints (Lan et al., 2023). Classical distribution alignment ideas also persist. Statistical colour transfer aligns channel wise distributions across images (Reinhard et al., 2002), and related illumination normalisation and colour constancy approaches estimate global or local corrections from image statistics (Buchsbaum, 1980; Finlayson & Trezzi, 2004; Van De Weijer et al., 2007). These ideas anticipate distribution matching objectives used in unpaired translation frameworks such as CycleGAN (Zhu, Park, et al., 2017). More broadly, modern day and night translation extends classical principles with explicit constraints that target semantics and stability, including semantic consistency and temporal coherence, addressing failure modes that classical methods cannot reliably control under severe illumination change (Isola et al., 2017; Schutera et al., 2020; T.-C. Wang, Liu, Zhu, Liu, et al., 2018).

In summary, classic illumination and colour normalisation methods explain important aspects of the day and night appearance gap, but their limitations under extreme degradation motivate learning based translation with constraints designed to preserve scene structure, object integrity, and temporal stability.

3 Scope, terminology and survey protocol

3.1 Scope and terminology

We survey methods whose primary objective is IDT, for outdoor scenes in RGB images and RGB video. IDT encompasses D2N and N2D translation and is motivated by downstream outdoor vision tasks including detection, segmentation, localisation, and tracking.

In this survey, translation means mapping an input from one illumination domain to another while preserving scene geometry and scene semantics. For video, translation also requires temporal coherence, meaning that static scene content should remain stable across frames. A representative example is translating night driving imagery into a daytime appearance so that a perception pipeline developed for daytime data fails less often under illumination shift.

We distinguish translation from enhancement and from domain adaptation. Enhancement improves visibility or perceptual quality within the same illumination domain, for example, brightening, denoising, or contrast correction of night images, and it does not target a distinct daytime domain. Domain adaptation aims to improve the robustness of a task model across domains, for example, by feature alignment, self training methods, or related representation learning objectives, and it does not require an explicit translated image or video as the final output.

To keep the scope precise, we treat enhancement and domain adaptation as context unless they are directly used to support IDT or to contextualise evaluation practice, and we do not include them in the main method set. We also apply explicit exclusions. We exclude cross modality translation, such as thermal to visible or infrared to visible, and related sensor translation settings. We exclude generic image to image translation methods that appear only as baselines and are not designed for illumination shift. We exclude pure enhancement methods whose objective is to improve night imagery without targeting the day as a distinct domain. We exclude domain adaptation methods that do not include explicit image or video translation or that do not evaluate the translation output.

3.2 Literature review and inclusion criteria

To define the set of works reviewed beyond the introduction, we followed a two stage literature collection strategy that combined keyword based retrieval with citation snowballing. We used Google Scholar for discovery and queried combinations of D2N and N2D translation terms with modelling phrases, including *day to night image conversion*, *night to day image conversion*, *image conversion*, *day to night image conversion using CNN*, *image to image translation*, *illumination translation*, *time of day translation*, *GAN based translation*, and *CNN based conversion*. The queries were iteratively refined by inspecting the results and adding synonym terms encountered in relevant papers. To reduce the risk of missing influential or less easily discoverable work, we then applied backward and forward citation snowballing on the resulting seed set by screening reference lists and papers that cite the seed papers. We repeated this process until additional iterations yielded few or no new relevant papers.

The main taxonomy and comparative analysis are based on papers that address the translation of D2N or N2D for outdoor scenes and provide sufficient methodological or empirical detail for comparison. We include a paper in this main set when it satisfies at least two conditions. The paper states an explicit objective for the D2N or N2D translation. The paper is motivated by outdoor driving or surveillance settings, or it uses datasets and evaluation protocols drawn from those settings. The method design is tailored to illumination shift, for example through physics motivated priors, semantic constraints, degradation disentanglement, or temporal coherence mechanisms. The paper reports on empirical evaluation of day and night datasets or assesses impact on downstream outdoor vision tasks. Generic image to image translation methods that are referenced only as baselines are not counted as part of the main set.

Applying this search and screening procedure yields a main set of 30 IDT methods for detailed analysis, where the count is an outcome of the process rather than an additional selection stage. We treat a method as a distinct entry when it introduces a materially different modelling assumption, constraint, or training signal that maps to our taxonomy, and we discuss closely related variants together to avoid counting minor revisions multiple times. For datasets, we report 22 public benchmarks for outdoor RGB imagery and video that are used in the included IDT papers as evaluation reference points. Although the procedure is systematic, omissions remain possible due to the breadth of related work, differences in terminology between communities, and new releases that appear after the search snapshot.

3.3 Artefact audit protocol

We also conduct an artefact availability audit for representative methods covered in this survey, and we define the audit protocol in subsection 8.1.

4 A constraint-centric taxonomy for IDT

IDT between day and night exhibits a severe appearance gap. Illumination changes are spatially non-uniform, local light sources introduce saturation and glare, sensor noise increases, and task-relevant cues can become weak or partially invisible. Under these conditions, objectives based primarily on distribution matching can fail by distorting structure, removing objects, or generating visually plausible outputs that violate scene semantics or scene geometry. These risks are particularly consequential in outdoor applications, such as automated driving and surveillance, where hallucinated content or missing objects can mislead downstream perception systems.

To make the literature easier to interpret, we organise methods using a constraint-centric taxonomy. Rather than grouping papers only by architecture, we group them by the explicit constraints and priors introduced to keep IDT faithful under extreme illumination shift. This perspective clarifies (i) which failure modes a method targets (e.g. semantic drift, hallucinated structure, temporal instability) and (ii) which evaluations are needed to substantiate claimed robustness.

4.1 Taxonomy dimensions

We describe methods along three complementary dimensions: the supervision available during training, the data modality at inference time, and the constraint families used to regularise the mapping under extreme illumination shift.

4.1.1 Supervision regime

The methods differ in the supervision used to learn the IDT mappings. Paired supervision provides pixel-level guidance when reliable alignment exists and can improve the preservation of fine structures (Isola et al., 2017; Punnappurath, Abuolaim, Abdelhamed, Levinshtein, & Brown, 2022). However, in outdoor settings, true day–night alignment is difficult due to dynamic objects, exposure variation, and viewpoint drift, and pseudo-pairs can introduce systematic bias. Unpaired supervision is therefore common because it is easier to scale, using adversarial objectives and consistency constraints to align domains without correspondence (M.-Y. Liu, Breuel, & Kautz, 2017; Yi, Zhang, Tan, & Gong, 2017; Zhu, Park, et al., 2017). Since distribution-level alignment does not guarantee semantic correctness under severe illumination changes, many works incorporate weak or auxiliary supervision—such as semantic labels, geometric cues, pseudo-pairs, or teacher networks—to anchor translation to meaningful structure (Bang et al., 2024; Bhattacharjee et al., 2020).

4.1.2 Data modality

Most early work focusses on still images, where each input is processed independently. In outdoor deployment, however, video IDT is often required, and temporal coherence becomes a first-order requirement: the translation must preserve scene semantics and geometry not only within each frame but also consistently across time. Video methods therefore introduce explicit temporal objectives, recurrent components, temporal discriminators, or motion-guided constraints (e.g. optical-flow-based consistency) to reduce flicker and identity drift across frames (Chen, Pan, Yao, Tian, & Mei, 2019; T.-C. Wang, Liu, Zhu, Liu, et al., 2018). This distinction matters because frame-wise translators can appear plausible on a per-frame basis while failing under temporal evaluation, with direct consequences for downstream tasks such as tracking and localisation.

4.1.3 Constraint and prior families

Across supervision regimes and data modalities, the main differentiator is the set of constraints and priors used to regularise IDT under severe ambiguity and partial observability.

- **Cycle and identity constraints** promote content preservation by encouraging invertibility between domains and discouraging unnecessary appearance changes; they form the backbone of many unpaired approaches (Zhu, Park, et al., 2017).
- **Semantic and instance constraints** preserve object identity and scene layout by coupling translation to task signals (e.g. segmentation or detection) or by enforcing consistency through pretrained task networks (Bhattacharjee et al., 2020; Shiotsuka et al., 2022).
- **Physics-informed priors** encode illumination and degradation structure explicitly—e.g. via decomposition, image formation models, or invariants—reducing reliance on purely statistical alignment (Y.-J. Lee, Go, Lee, Son, & Lee, 2025).
- **Contrastive and correspondence constraints** enforce local consistency through

patch, feature, or correspondence-level matching between related regions, which can improve structural preservation and mitigate collapse in ambiguous areas (Lan et al., 2023; Park, Efros, Zhang, & Zhu, 2020).

- **Temporal constraints** enforce frame-to-frame coherence in video IDT, mitigating flicker, colour instability, and identity drift that degrade tracking and localisation (T.-C. Wang, Liu, Zhu, Liu, et al., 2018).

Table 2 provides a crosswalk from methods to constraints for IDT. Each row corresponds to a representative method. The *Sup.* column indicates the supervision regime, and the *Gap* column lists the primary domain gap factors addressed using codes I (illumination), G (glare), N (noise), W (weather), and M (motion or video). The constraint and prior columns indicate which constraint families are central to the method, using ticks to mark Cycle or Identity, Semantic or Instance, Correspondence, Physics, and Temporal constraints. The final columns summarise the most defensible evaluation evidence emphasised in the original work, using P (perceptual and distributional), S (semantic and structural), D (downstream task utility) and T (temporal stability).

4.2 Failure modes and constraint selection

IDT between day and night exhibits failure modes—including semantic distortions and temporal flicker—that are not reliably reflected by perceptual quality measures alone (Blau & Michaeli, 2018; Z. Jia et al., 2021). Characterising these failures is therefore useful both for selecting constraints and priors and for interpreting reported results.

Unpaired translators can satisfy cycle consistency while still altering object identity or scene geometry in severely under-illuminated regions, a behaviour commonly described as *semantic drift* (Bhattacharjee et al., 2020; Shiotsuka et al., 2022). In regions with limited evidence—particularly in N2D translation, translation—generative models can introduce a visually plausible but incorrect structure, reflecting ambiguity in conditional mapping and objectives that prioritise perceptual plausibility (Blau & Michaeli, 2018). Global normalisation and global losses can also suppress weak signals from small or distant objects, motivating patch- or correspondence-level constraints that preserve local structure (Lan et al., 2023; Park et al., 2020). In the video setting, independent per-frame translation frequently produces flicker and identity drift over time, motivating explicit temporal regularisation when translation is applied to streams (Chen et al., 2019; T.-C. Wang, Liu, Zhu, Liu, et al., 2018).

In practice, no single constraint is sufficient in all night conditions. Effective systems balance (i) content preservation, (ii) semantic anchoring, (iii) robustness to illumination degradation, and—when operating in video, (iv) temporal stability.

4.3 Checklist for assessing new methods

When assessing a D2N translation approach, it is useful to ask four questions. Which failure mode is targeted, which constraint family addresses it, what evidence is provided beyond visual examples, and whether the method is reproducible in the sense that training and evaluation details are sufficiently available to support verification and fair comparison.

5 Learning based translation methods

This section reviews learning-based approaches for IDT of outdoor RGB images and videos, encompassing N2D and D2N translation. We organise the literature primarily by the supervision regime and then by the constraints used to preserve scene semantics, scene geometry, and—when operating on video—temporal coherence under severe illumination change. Throughout, we emphasise design choices that are particularly relevant to driving and surveillance, including semantic and instance preservation, robustness to multi-modal night illumination, task-driven objectives, physics-informed priors, and temporal consistency.

5.1 Paired translation with supervised conditional models

Paired translation learns a direct mapping G between illumination domains using aligned training pairs (e.g. N2D pairs). Conditional generative adversarial networks provide a canonical formulation in this setting. Pix2Pix uses a conditional adversarial objective together with an ℓ_1 reconstruction term to encourage realism while preserving pixel-level fidelity (Isola et al., 2017). When alignment is reliable, supervised training provides strong guidance and can preserve fine structures more consistently than unpaired alternatives.

In practice, the main limitation is data acquisition. Collecting truly aligned day - night pairs in unconstrained outdoor environments is difficult because dynamic objects, exposure variation, weather, and small viewpoint drift break pixel correspondence. As a result, supervised pipelines frequently rely on pseudo-pairs or approximate alignment, for example, by synthesising one domain from the other and then training a supervised model on the resulting pairs (Punnappurath et al., 2022). These strategies can improve scalability, but they introduce a risk of *bias inheritance*, where artefacts or colour statistics introduced during pseudo-pair generation are learnt and sometimes amplified by the supervised translator.

5.2 Unpaired translation with adversarial alignment and consistency

Unpaired translation learns mappings between collections of day- and night images without explicit correspondence. CycleGAN couples two generators through adversarial alignment and a cycle-consistency objective that encourages translated images to map back to their inputs, and often includes an identity term to discourage undesirable colour shifts (Zhu, Park, et al., 2017). Closely related formulations include dual learning approaches (Yi et al., 2017) and shared-latent-space models such as UNIT, which combine variational encoding with adversarial learning under the assumption that corresponding images share a common representation (M.-Y. Liu et al., 2017). These methods remain widely used as baselines because they eliminate the need for paired data.

However, for IDT, the distribution-level alignment and the cycle consistency are not sufficient to guarantee semantic or geometric faithfulness. Mixed artificial lighting, saturated highlights, sensor noise, and deep shadows can allow cycle constraints to be satisfied while object identity, boundaries, or even the presence of small but critical road users changes during translation. This motivates additional constraints that explicitly preserve semantics and structure.

5.3 Constraints for semantic and structural preservation

Outdoor vision applications require stable preservation of roads, lane markings, traffic signs, pedestrians, and vehicles. Semantically guided translation introduces explicit task signals-typically via segmentation guidance or semantic-consistency losses-to anchor the mapping to a meaningful structure (Bang et al., 2024; Shiotsuka et al., 2022). Instance-aware approaches further emphasise object-level integrity. DUNIT integrates an object detector into the translation pipeline and enforces instance-level consistency, improving object preservation, and supporting downstream detection under domain shift (Bhattacharjee et al., 2020). In this context, these constraints primarily address *semantic drift*, where visually plausible translations distort object boundaries or alter object identity.

A related strategy focusses on local correspondence rather than global distribution matching. Contrastive and correspondence-based constraints encourage patch-level consistency and can reduce structural distortions in ambiguous regions. Disentanglement-based models pursue a complementary objective by separating domain-invariant content from domain-specific illumination or style, supporting controllability, and reducing collapse toward an average appearance. For example, DiCo combines disentanglement with contrastive learning and introduces an explicit prior to improve stability under challenging surveillance illumination (Lan et al., 2023). In practice, these approaches are best interpreted as robustness mechanisms rather than diversity mechanisms alone, since they are designed to preserve structural signals that matter for downstream outdoor tasks.

5.4 Task driven translation for localisation and retrieval

A distinct family of approaches treats IDT as a means of improving a downstream task rather than as an end goal. For example, retrieval-based localisation under D2N variation benefits from outputs that preserve matchable structure and stable correspondences, not necessarily photorealistic appearance. ToDayGAN adapts unpaired translation with task-motivated discriminators and constraints that emphasise useful cues for retrieval and localisation under large appearance gaps (Anoosheh et al., 2019). Related work shows that translation (and, in some pipelines, enhancement used for data generation or pre-processing) can expose localisation systems to difficult night conditions, with evaluation centred on retrieval recall and pose accuracy rather than image similarity (Mohwald, Jenicek, & Chum, 2023). For this family, perceptual metrics can be misleading; task-specific measures such as match statistics, Recall@K, and pose error are essential for a meaningful comparison.

5.5 Physics informed and degradation disentangling models

Physics-informed approaches treat IDT as more than generic appearance transfer by explicitly modelling night-specific degradations such as illumination deficiency, sensor noise, and saturated highlights. Many methods adopt a decomposition principle, separating the input into interpretable factors (e.g. illumination and reflectance) or regions (e.g. glare versus under-exposed areas), applying factor-specific constraints, and recombining a daytime-like output while enforcing content fidelity. Recent work on disentangling illumination degradation introduces degradation-sensitive objectives, including contrastive terms, to improve robustness under extreme night conditions (Lan et al., 2024). The appeal of this family lies in its inductive bias: generalisation is supported by assumptions about image formation and illumination behaviour rather than relying solely on distribution matching.

5.6 Temporal constraints (overview)

Many deployment settings require video illumination-domain translation (IDT), where frame-wise translation commonly produces flicker and temporal identity drift. Video methods therefore introduce temporal constraints (e.g. motion-compensated consistency, spatio-temporal objectives, recurrent/3D architectures, or feature-space coherence) to stabilise outputs over time.

Because these mechanisms, their failure cases specific to the night, and the evaluation specific to the video deserve a focused treatment, we review them in section 6.

5.7 Architectural refinements and practical variants

Beyond baseline unpaired formulations, many works propose architectural refinements to improve stability and structure preservation under severe illumination shift. Examples include multi-scale designs, attention mechanisms, and structural constraints such as edge or gradient consistency, as well as models that explicitly target discriminative cues under day-night variation (H. Liu, Cheng, & Ye, 2024; Taufiq & Rahadianti, 2025; Torbunov et al., 2023). Although such refinements can improve visual quality, their impact should be assessed together with semantic preservation and task utility, particularly in safety-relevant applications.

6 Video translation and temporal consistency

Video IDT extends image-based translation from isolated frames to sequences, where preserving temporal coherence is as important as preserving per-frame semantics and geometry. Applying an image translator independently to each frame commonly produces flicker in colour and tone, edge instability, and identity drift for small objects. These artefacts reduce operator trust and can materially affect downstream modules such as tracking, mapping, and motion forecasting.

Most effective pipelines therefore introduce constraints that explicitly couple adjacent frames and stabilise appearance over time. A common formulation is flow-guided temporal regularisation, where translated outputs or intermediate features are warped between consecutive frames and deviations are penalised, typically with occlusion handling to avoid over-constraining disoccluded regions (Lai et al., 2018; Ruder, Dosovitskiy, & Brox, 2016; T.-C. Wang, Liu, Zhu, Liu, et al., 2018). Motion-aware designs further incorporate temporal structure into the learning objective, for example, through discriminators that operate on short frame windows so that appearance and temporal dynamics are judged jointly (Chen et al., 2019; T.-C. Wang, Liu, Zhu, Liu, et al., 2018). Recurrent generators and three-dimensional architectures provide an alternative by propagating information over time to reduce frame variance and stabilise repeated structures (T.-C. Wang, Liu, Zhu, Liu, et al., 2018; Wei, Zhu, Feng, & Su, 2018). More recently, feature-space coherence has been proposed as a less brittle alternative to pixel-space constraints under motion blur and illumination variation, enforcing temporal stability in learnt representations (Yang, Zhou, Liu, & Loy, 2024).

Representative mechanisms can be summarised as follows.

- Optical-flow-guided warping losses that penalise inconsistencies after motion compensation (Lai et al., 2018; T.-C. Wang, Liu, Zhu, Liu, et al., 2018).
- Spatio-temporal discriminators and motion-aware objectives that encourage realistic short-term dynamics (Chen et al., 2019; T.-C. Wang, Liu, Zhu, Liu, et al., 2018).
- Recurrent and three-dimensional designs that propagate context and reduce frame-to-frame variance (Wei et al., 2018).
- Feature-space temporal coherence objectives that stabilise intermediate representations (Yang et al., 2024).

Nighttime video IDT remains substantially harder than its daytime counterpart because illumination dynamics after dark violate assumptions underlying many temporal constraints. Sudden exposure changes, moving headlights, specular reflections, and localised light sources undermine brightness constancy and smooth temporal variation, allowing appearance drift to accumulate over longer sequences even when short-term consistency is satisfactory (T.-C. Wang, Liu, Zhu, Liu, et al., 2018; Wei et al., 2018). Flow-guided constraints are particularly fragile at night because low texture, motion blur, and sensor noise degrade flow reliability, and incorrect warps may be enforced as if they were ground-truth correspondences (Lai et al., 2018; Ruder et al., 2016). Spatio-temporal discriminators can improve global temporal realism, but they do not necessarily protect the identity of fine-grained objects, especially small or distant objects that contribute weakly to the adversarial signal (Chen et al., 2019). Feature-space coherence mitigates some of these issues, but its effectiveness depends on the robustness of the chosen representations under severe illumination variation (Yang et al., 2024).

Long-horizon stability introduces additional requirements beyond the adjacent-frame coupling. Many pipelines achieve short-term coherence by conditioning on neighbouring frames, yet degrade over longer sequences when objects leave and reenter the field of view or when lighting changes abruptly. Mechanisms for longer-term consistency have therefore been explored in video synthesis and translation (Mallya, Wang, Sapra, & Liu, 2020; Rivoir et al., 2021), alongside practical acceleration strategies that reduce latency while maintaining temporal stability (Zhuo, Wang, Li, Wu, & Liu, 2022). The application context further modulates the difficulty: moving-camera driving video introduces parallax, motion blur, and frequent dynamic objects, while fixed-camera surveillance reduces viewpoint variation but remains challenging under mixed illumination events, such as headlights entering and exiting the scene.

From an evaluation perspective, video IDT should be assessed using temporal metrics in addition to per-frame realism. Common choices include motion-compensated consistency measures such as warping error (Lai et al., 2018) and sequence-level stability criteria that quantify drift over time (Wei et al., 2018). For outdoor vision, it is also important to report the impact on downstream tracking and localisation, since visually smooth outputs may still suppress small but safety-critical objects or degrade matchability (T.-C. Wang, Liu, Zhu, Liu, et al., 2018; Wei et al., 2018).

7 Datasets and evaluation for outdoor vision

Datasets and evaluation protocols largely determine the conclusions drawn in IDT between day and night. In outdoor vision, the domain gap is driven not only by illumination, but also by weather, seasonality, exposure control, sensor noise, and the motion and viewpoint dynamics of the sensing platform. This section summarises dataset properties that materially affect training and empirical claims, reviews representative benchmarks used in the literature, and proposes an evaluation protocol that links translation behaviour to downstream outdoor-vision utility.

7.1 Datasets

Outdoor D2N datasets differ in ways that strongly influence both what a model can learn and what the evaluation results mean. Moving-platform data, typical in driving, introduce parallax, motion blur, rolling-shutter artefacts, and frequent dynamic objects, while fixed-camera surveillance reduces viewpoint variation, but can include sharp illumination transients such as headlights entering and leaving the scene (Neumann et al., 2018). The availability of correspondence is another key distinction: some datasets provide paired or pseudo-paired structure, while many are unpaired collections. Cross-time correspondences are particularly valuable for evaluation because they help control content changes when comparing day and night observations (Sakaridis et al., 2019, 2025). Adverse conditions beyond illumination also matter. Rain, fog, and snow interact with night lighting and can dominate the appearance gap, so datasets that explicitly include these conditions reduce the risk that models are tuned only to clear night scenes (Sakaridis et al., 2025). Finally, the type and reliability of the annotations determine what can be evaluated. Pixel-level labels, instance masks, and detection boxes enable task-based evaluation, but nighttime introduces genuinely ambiguous regions, motivating uncertainty-aware protocols when available (Sakaridis et al., 2019, 2025). Table 4 summarises these characteristics of the data set for representative benchmarks and, crucially, indicates which evaluation evidence is well supported given the available correspondence structure, modalities, and annotations.

No single benchmark covers all deployment conditions, so the literature relies on complementary datasets. For driving and dense prediction, Dark Zurich provides day, twilight and night imagery with correspondences and a labelled nighttime benchmark designed for uncertainty-aware evaluation (Sakaridis et al., 2019). ACDC extends this setting to multiple adverse conditions, including night, and provides correspondence structure and panoptic-style annotations (Sakaridis et al., 2025). Nighttime Driving is smaller, but widely used for nighttime segmentation comparisons (Dai & Gool, 2018), and NightCity provides a larger labelled set that highlights exposure effects in urban scenes (Tan et al., 2021). For broader driving pipelines, large suites such as BDD100K include night content in realistic operating conditions (Yu et al., 2020), while datasets such as nuScenes and Waymo are often used when translation is evaluated indirectly through downstream detection and tracking in multi-sensor settings (Caesar et al., 2020a; Sun et al., 2020).

For localisation under changing conditions, Aachen Day–Night and RobotCar Seasons are commonly used benchmarks for retrieval and pose estimation on large illumination and condition shifts (Sattler et al., 2018). The Oxford RobotCar dataset provides repeated traversals of the same route between seasons and lighting, supporting long-term cross-condition evaluation (Maddern, Pascoe, Linegar, & Newman, 2017). For detection at night, NightOwls is a dedicated pedestrian benchmark with dense annotations in long video sequences (Neumann et al., 2018), and ExDark provides object annotations in low-light imagery that is frequently used in recognition and detection studies, including translation-based pre-processing (Loh & Chan, 2019). When the setting involves multi-sensor fusion, aligned visible-thermal benchmarks such as KAIST and LLVIP become relevant (Hwang, Park, Kim, Choi, & So Kweon, 2015; X. Jia, Zhu, Li, Tang, & Zhou, 2021), and automotive thermal datasets such as FLIR ADAS are used when translation is evaluated as part of a thermal perception pipeline (Teledyne FLIR LLC, 2018).

7.2 Evaluation protocol and common pitfalls

A recurring flaw in the translation of D2N and N2D is *the metric mismatch*, that is, commonly reported image metrics are weak predictors of outdoor vision utility. Pixel metrics such as PSNR and SSIM assume accurate spatial correspondence and are therefore meaningful only under truly aligned pairs. In many D2N benchmarks, correspondence is approximate, synthetic, or absent, making pixel based scores difficult to interpret and easy to over claim (Sakaridis et al., 2019, 2025). When ground truth is not available, distributional and perceptual measures such as FID and LPIPS are often reported (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017; R. Zhang et al., 2018). However, these metrics can miss critical local safety failures, such as small object removal, lane boundary deformation, and traffic sign hallucination. No reference image quality metrics such as NIQE and BRISQUE are also problematic because their underlying natural image statistics are systematically violated in nighttime imagery (Mittal, Moorthy, & Bovik, 2012; Mittal, Soundararajan, & Bovik, 2013).

Table 1 summarises the four evidence categories used throughout this survey. We annotate each paper with one or more evidence codes: P for perceptual and distributional similarity, S for semantic and structural preservation, D for downstream task utility and T for temporal stability in video settings. When reporting results, authors should state which components are supported by the dataset and which are evaluated in the paper, and avoid interpreting a single component as sufficient evidence of utility. Pixel based metrics should be reported only when pairing is truly aligned and should not be treated as a proxy for perceptual quality under approximate correspondence. When claims concern robustness for detection, segmentation, tracking, localisation, or retrieval, we treat downstream evidence on real nighttime data as essential. For video translation, temporal evidence is expected in addition to per frame reporting.

7.3 Integration driven evaluation

The practical value of IDT depends on how it is used in an outdoor-vision pipeline. In some settings, it acts as a pre-processing normaliser, in others it serves as data augmentation, and in others it is embedded in self-training or curriculum adaptation. These integration choices determine which failure modes matter and which evidence is meaningful.

For system testing, translation can generate controlled nighttime variants of daytime driving scenes, allowing consistency checks through metamorphic relations between model output on an original input and its systematically transformed counterpart (Tian, Pei, Jana, & Ray, 2018; M. Zhang, Zhang, Zhang, Liu, & Khurshid, 2018). In this setting, the key requirement is to specify the validity conditions for the transformation and to report the behavioural consistency criteria. For detection-orientated use, N2D translation can be applied online as a front-end normalisation step when detectors are trained predominantly on daytime data (Schutera et al., 2020). Here, latency and temporal stability matter, and evaluation should report detector performance across multiple architectures and datasets, with attention to small-object preservation. Instance-aware translation is particularly relevant because it directly targets the integrity of objects (Bhattacharjee et al., 2020). For semantic segmentation and adaptation, translation is often coupled with correspondence signals, curriculum, or self-training to transfer daytime supervision to nighttime imagery (F. Huang, Yao, & Zhou, 2023; Sakaridis et al., 2019; Wu, Wu, Guo, Ju, & Wang, 2021). In this setting, semantic preservation evidence and benchmarking on established nighttime protocols are central. For localisation and retrieval, the objective is matchability rather than photorealism, so evaluation should prioritise retrieval recall, keypoint match statistics, and pose error (Anoosheh et al., 2019).

Across these integration settings, a consistent implication is that realism metrics alone are insufficient. The evidence should match the operational role of translation and expose the failure modes most likely to affect downstream outdoor-vision performance.

Table 1: Four component evaluation protocol for IDT.

Code	Component	What to report in practice
P	Perceptual and distributional similarity	Use coarse realism checks when ground truth is unavailable. Report FID, optionally add KID; when paired or approximately paired evaluation exists, report LPIPS. Avoid pixel metrics unless alignment is truly valid.
S	Semantic and structural preservation	When labels or correspondences exist, evaluate preservation of semantics and structure. For segmentation with ambiguous nighttime regions, use uncertainty aware protocols when available. For localisation, prioritise matchability evidence such as keypoint statistics and pose error.
D	Downstream task utility	When claims concern robustness or deployment benefit, report task performance on real nighttime data, for example detection, segmentation, tracking, or localisation metrics.
T	Temporal stability for video	When translation is applied to video, evaluate temporal coherence using warping based consistency measures and sequence level stability criteria, alongside per frame scores and relevant sequence based downstream tasks.

8 Artefact availability review

Artefact availability matters because IDT papers often report gains under severe illumination changes, but empirical results can be difficult to reproduce without a complete research compendium, including the codebase, data processing pipeline, checkpoints, and experimental settings (Gundersen & Kjensmo, 2018; Haibe-Kains et al., 2020; Pineau et al., 2021). This concern is particularly acute for generative models, where outcomes can vary with implementation details, optimisation schedules, and evaluation pipelines (Gundersen, Coakley, Kirkpatrick, & Gil, 2022).

8.1 Audit protocol

We audited publicly available artefacts for representative methods referenced in this survey using official project pages and author maintained repositories where possible. In addition to official project pages and author maintained repositories, we also checked links provided in the paper and supplementary material, as well as any archival releases referenced from those sources. The audit reflects a snapshot as of **December 19, 2025** , and may change as materials are released or updated. This audit is a snapshot of the availability of public artefacts. We verified public accessibility and documentation of the reported materials, but we did not retrain models, execute training pipelines, or run evaluation scripts. Availability therefore reflects whether an artefact is publicly accessible and documented to an operational standard, not whether it reproduces reported numbers when executed.

For each method, we record five availability signals using operational criteria. The **Code** is marked as available only when a publicly accessible repository or archived release provides training and inference code sufficient to run the method from start to finish from input data to translated outputs. **Weights** are marked as available only when pretrained checkpoints are provided for at least one reported setting. **Data** is marked as available only when working links and acquisition instructions are provided for the datasets used in the main experiments, including any access procedures. The **Recipe** is marked as available only when the release includes configuration or hyperparameter files, an environment specification, and explicit commands or scripts for training and evaluation, together with dataset preprocessing steps and data set splits used in the paper. The **license** is marked as available only when the release includes explicit licence terms for code and, where specified, separate terms for pretrained weights.

We prioritise artefacts released by the original authors or their institutions. Third party reimplementations may be informative engineering references, but they are not treated as definitive evidence of reproducibility unless explicitly validated against the original results.

Table 3 reports on the artefact availability matrix for the representative methods covered in this survey. It indicates, per method, whether code, pretrained weights, dataset access details, a runnable training and evaluation recipe, and explicit licence terms were publicly available at the time of the audit. For readability, we use ✓ to indicate available and a dash to indicate not identified.

8.2 Artefact matrix and implications

Table 3 summarises the availability of artefacts for selected methods. The table is intended as a transparency aid rather than a judgement of scientific merit. A method can be technically strong even if artefacts are not publicly released, but the burden of verification then shifts to the reader.

The following two implications are given. First, comparisons should be conditioned on the reproducibility tier. When weights or a complete recipe are missing, reported gains can be difficult to validate under a common protocol, and the cost of reproduction can exceed the apparent margin of improvement. Second, licencing should be treated as a first-class deployment constraint: in safety-relevant outdoor systems, unclear licencing terms for code or pretrained weights can block adoption even when an approach is technically promising.

We therefore encourage future work to release, at minimum, a versioned repository with training and evaluation scripts, the exact dataset splits used in the article, and explicit licences for code and pretrained weights (Haibe-Kains et al., 2020; Pineau et al., 2021).

9 Open problems and future directions

Despite substantial progress in D2N image and video translation, several challenges remain that limit robustness, generalisation, and safe deployment in real-world outdoor vision systems. To make these challenges actionable, we state them in terms of dataset and protocol requirements: what future benchmarks should contain and what evaluation should report so that claims become falsifiable and comparable.

9.1 Heterogeneous night conditions and semantic faithfulness

Nighttime outdoor scenes exhibit strong spatial heterogeneity, combining under-exposed regions with saturated highlights from headlights, street lamps, and reflective surfaces. Methods relying on global appearance shifts often fail in this regime, for example, by over-brightening dark areas or suppressing highlight structure and colour information (Jobson et al., 1997; Neumann et al., 2018). The problem is compounded by ambiguity: a translated image can appear realistic while altering traffic signs, removing pedestrians, or hallucinating lane markings, which is especially risky in safety-critical settings (Blau & Michaeli, 2018; R. Zhang et al., 2018). Region-aware translation and physics-guided decomposition aim to address mixed illumination by separating illumination components and applying different constraints to dark and glare-dominated regions (Lan et al., 2024; Y.-J. Lee et al., 2025), while semantic and instance constraints aim to anchor translation to task-relevant structure (Bhattacharjee et al., 2020; Shiotsuka et al., 2022). However, existing benchmarks rarely stress-test these conditions in a controlled manner.

Future datasets should therefore include explicit subsets with mixed lighting events (headlight glare, specular reflections, deep shadows) and provide region-level evaluation hooks. At minimum, these hooks can be implemented via reproducible proxies such as saturation/highlight masks and under-exposure masks; for a subset, human-verified region annotations further improve diagnostic value. When label supervision exists, benchmarks should include annotations that enable semantic or instance evaluation on critical classes, and when scenes contain genuinely ambiguous pixels or instances, uncertainty-aware labelling and evaluation should be adopted rather than excluding difficult regions without disclosure (Sakaridis et al., 2019, 2025).

Evaluation should be reported stratified by degradation severity (e.g. saturation fraction and under-exposure fraction) and by object size bins, because many safety-relevant errors concentrate in small objects and low-evidence regions. In addition to perceptual metrics, protocols should require explicit semantic-faithfulness evidence such as task-network consistency (e.g. segmentation/detection consistency before and after translation), object retention measures for critical classes, and explicit accounting of both object removal and hallucination rather than only qualitative examples. Reporting should include region-wise summaries and representative failure cases conditioned on region masks, so that improvements cannot be driven solely by global image-level scores.

9.2 Generalisation across locations, weather, and sensors

Many translation models are overfitting to specific camera pipelines, cities, or lighting infrastructures. Differences in sensor response, noise characteristics, and illumination spectra can degrade performance when models are applied outside of their training domain (J. Liu et al., 2021; Sakaridis et al., 2025). Adverse weather further compounds the appearance gap and interacts with night illumination in ways that are not captured by clear-night benchmarks (Sakaridis et al., 2025).

Benchmarks should explicitly expose domain factors (at least location and sensor/camera pipeline; ideally weather) with enough samples per factor level to support stratified evaluation. Standard splits should include held-out domain settings (e.g. an unseen city or unseen sensor pipeline) and, where possible, combined adverse conditions (e.g. night+rain/fog/snow) to prevent tuning to clear-night statistics.

Claims of robustness should be supported by cross-domain evidence: reporting should include held-out-domain performance, and, when multiple factors exist, a factor-wise performance matrix rather than a single average. Improvements in a single benchmark should be described as in-domain gains unless they are validated across datasets or explicitly held-out domains.

9.3 Temporal stability for video IDT

For surveillance and automated driving, temporal instability is unacceptable. Flicker, colour oscillation, and identity drift undermine operator trust and can degrade tracking, localisation, and mapping pipelines (T.-C. Wang, Liu, Zhu, Liu, et al., 2018; Wei et al., 2018). Although recent video translation methods introduce temporal constraints, evaluation still often prioritises per-frame realism. Methods relying on optical flow should explicitly analyse failure cases in low-texture and noisy nighttime conditions, where flow estimates are unreliable (Lai et al., 2018; Ruder et al., 2016).

Video benchmarks should include sequences with night-specific temporal stressors (headlights entering and exiting, exposure shifts, specular sweeps) and sufficient horizon length to measure drift, not only short clips. Evaluation should treat temporal metrics as first-class outputs alongside per-frame realism: motion-compensated consistency measures and sequence-level stability indicators should be reported, and long-horizon behaviour should be summarised explicitly (e.g. stability as a function of time or over event segments containing illumination transients). For outdoor vision relevance, reporting should include downstream sequence-based evaluation (e.g. video tracking/localisation) rather than only frame-wise snapshots (T.-C. Wang, Liu, Zhu, Liu, et al., 2018; Wei et al., 2018). For flow-based methods, reports should include sensitivity analyses in flow failure modes typical of nighttime video (Lai et al., 2018; Ruder et al., 2016).

To support fair comparison and long-term value in archival venues, reporting should standardise dataset splits and preprocessing pipelines, training schedules and compute, and ablations isolating the effect of key constraint families (Cyc/Id, Sem/Inst, Corresp, Phys, Temp). When claims concern downstream utility, evaluation should include more than one perception model to reduce model-specific bias. Finally, artefact availability should be reported with versioning and explicit licences for both code and pretrained weights.

Table 2: Representative crosswalk from methods to constraints for IDT. Gap codes: I: illumination, G: glare, N: noise, W: weather, M: motion or video. Constraint codes: Cyc or Id: cycle and identity, Sem or Inst: semantic or instance or task network constraints, Corresp: contrastive or correspondence, Phys: physics or decomposition, Temp: temporal. Primary evaluation evidence codes follow Table 1: P: perceptual and distributional, S: semantic and structural, D: downstream task utility, T: temporal stability.

Method	Sup.	Gap	Constraints / priors					Primary eval.			
			Cyc/Id	Sem/Inst	Corresp	Phys	Temp	P	S	D	T
Pix2Pix (Isola et al., 2017)	Paired	I						✓	✓		
CycleGAN (Zhu, Park, et al., 2017)	Unpaired	I	✓					✓			
UNIT (M.-Y. Liu et al., 2017)	Unpaired	I	✓					✓			
CUT (Park et al., 2020)	Unpaired	I			✓			✓			
ToDayGAN (Anoosheh et al., 2019)	Unpaired/ task	I	✓	✓						✓	
DUNIT (Bhattacharjee et al., 2020)	Unpaired/ inst	I	✓	✓					✓	✓	
N2D-GAN (X. Li et al., 2022)	Unpaired/ weak	I	✓	✓				✓			
SGA-D2N (Bang et al., 2024)	Unpaired/ sem+geom	I	✓	✓				✓			
SPN2D-GAN (X. Li & Guo, 2023)	Unpaired/ sem	I	✓	✓					✓		
RefN2D-Guide GAN (Ning & Gong, 2023)	Unpaired/ ref+sem	I		✓	✓			✓	✓		
DiCo (Lan et al., 2023)	Unpaired	I			✓			✓	✓		
N2D3 (Lan et al., 2024)	Unpaired	I,G,N			✓	✓		✓	✓		
RLA-Training (Y.-J. Lee et al., 2025)	Unpaired+ paired/task	I	✓			✓				✓	
AU-GAN (Kwak et al., 2021)	Unpaired	I,W	✓					✓			
Daydriex (E. Lee & Kang, 2021)	Unpaired/ aux	I	✓		✓			✓		✓	
vid2vid (T.-C. Wang, Liu, Zhu, Liu, et al., 2018)	Paired/ (vid)	I,M					✓	✓			✓
MoCycleGAN (Chen et al., 2019)	Unpaired/ (vid)	I,M	✓				✓	✓			✓
UVCGAN (Torbunov et al., 2023)	Unpaired	I	✓					✓			
D2N ISP Synthesis (Punnappurath et al., 2022)	Paired/ (synth)	I,N				✓				✓	
Paired-N2D (Lakmal et al., 2024)	Paired/ (synth)	I						✓	✓		
DualGAN (Yi et al., 2017)	Unpaired	I	✓					✓			
CoMoGAN (Pizzati et al., 2021)	Unpaired	I				✓		✓			
ManiFest (Pizzati et al., 2022)	Few-shot/ unpaired	I						✓			
PITI (T. Wang et al., 2022)	Paired/ Unpaired	I						✓			
CycleGAN-Turbo (Parmar et al., 2024)	Unpaired	I,W						✓			
pix2pix-Turbo (Parmar et al., 2024)	Paired	I,W						✓			
Dark Side Augmentation (Mohwald et al., 2023)	Unpaired/ task	I		✓						✓	

Continued on next page

Table 2 (continued)

Method	Sup.	Gap	Constraints / priors					Primary eval.			
			Cyc/Id	Sem/Inst	Corresp	Phys	Temp	P	S	D	T
2PCNet (Kennerley et al., 2023)	Unpaired/ task	I,G,N		✓						✓	
Night-to-Day (Online) (Schutera et al., 2020)	Unpaired/ (vid)/task	I,M		✓			✓			✓	
ForkGAN (Z. Zheng et al., 2020)	Unpaired	I,W				✓		✓		✓	

Table 3: Artefact availability matrix for selected methods referenced in this survey, derived from publicly available repositories and project pages. "Recipe" denotes training and evaluation details sufficient for a faithful rerun. "Licence" denotes explicit licensing terms for code and, where applicable, pretrained weights. A blank entry indicates that the corresponding artefact was unavailable, not identified, or not clearly specified at the time of the audit and should be independently verified.

Paper / Method	**Focus**	**Code**	**Wts.**	**Data**	**Rec.**	**Lic.**
Pix2Pix (Isola et al., 2017)	Paired translation baseline	✓	✓	✓	✓	✓
CycleGAN (Zhu, Park, et al., 2017)	Unpaired translation baseline	✓	✓	✓	✓	✓
UNIT (M.-Y. Liu et al., 2017)	Shared-latent baseline	✓	-	✓	✓	-
ToDayGAN (Anoosheh et al., 2019)	N2D for localisation	✓	✓	✓	✓	✓
Daydriex (E. Lee & Kang, 2021)	N2D driving pipeline	-	-	-	-	-
DiCo Lan et al. (2023)	N2D for surveillance	-	-	-	-	-
N2D^3 (Lan et al., 2024)	Degradation disentanglement	-	-	✓	-	-
AU-GAN (Kwak et al., 2021)	Adverse-condition translation (incl. N2D)	✓	✓	✓	✓	-
DUNIT (Bhattacharjee et al., 2020)	Instance-aware translation	✓	-	✓	✓	-
SPN2D-GAN (X. Li & Guo, 2023)	Semantic-prior N2D	✓	-	✓	-	-
Paired-N2D (Lakmal et al., 2024)	Supervised N2D (synthetic pairs)	✓	-	✓	-	-
GMA-CycleGAN (Yang et al., 2023)	Thermal-to-visible translation	-	-	✓	-	-
IR2VI (S. Liu et al., 2018)	Thermal-to-visible (two-stage)	-	-	-	-	-
PITI (T. Wang et al., 2022)	Pretrained diffusion prior for translation	✓	✓	✓	✓	✓
CUT (Park et al., 2020)	Contrastive unpaired translation baseline	✓	-	✓	✓	✓
vid2vid (T.-C. Wang, Liu, Zhu, Liu, et al., 2018)	Video-to-video translation baseline	✓	✓	✓	✓	✓
UVCGAN (Torbunov et al., 2023)	UNet-ViT CycleGAN variant (strong unpaired baseline)	✓	✓	✓	✓	✓
CoMoGAN (Pizzati et al., 2021)	Continuous time-of-day translation (Day2Timelapse)	✓	✓	✓	✓	✓
Dark Side Augmentation (Mohwald et al., 2023)	D2N augmentation for retrieval / metric learning	✓	✓	✓	✓	✓
CycleGAN-Turbo / pix2pix-turbo (Parmar et al., 2024)	One-step translation using text-to-image backbones (incl. D2N, N2D)	✓	✓	✓	✓	✓
D2N ISP Synthesis (Punnappurath et al., 2022)	D2N synthesis (camera/ISP-aware)	✓	✓	✓	✓	✓
2PCNet (Kennerley et al., 2023)	D2N UDA object detection (NightAug, 2-phase consistency)	✓	✓	✓	✓	✓
traffic-pipeline (Alam, Parmar, et al., 2025)	Traffic video translation pipeline (N2D, D2N; clear-to-rainy)	✓	✓	✓	✓	✓
DualGAN (Yi et al., 2017)	Unpaired translation baseline (dual learning)	✓	✓	✓	✓	✓
EnlightenGAN (Jiang et al., 2021)	Low-light enhancement baseline (if included)	✓	✓	✓	✓	✓
SCC (Structure Consistency Constraint) (Guo et al., 2022)	Structure-preserving constraint for unpaired translation	✓	-	-	-	-
WKD (L. Zhang et al., 2022)	Efficient translation via distillation (wavelet KD)	✓	✓	✓	-	-
ManiFest (Pizzati et al., 2022)	Few-shot translation (includes D2N/day2twilight)	✓	✓	✓	✓	✓
SAVI2I (Mao et al., 2022)	Continuous multi-domain translation via signed attribute vectors	✓	✓	✓	✓	✓
InstaFormer (Kim et al., 2022)	Instance-aware I2I with Transformer	✓	-	✓	✓	-
InstaGAN (Mo et al., 2018)	Instance-aware I2I (object-centric translation)	✓	✓	✓	✓	✓

Continued on next page

Paper / Method	**Focus**	**Code**	**Wts.**	**Data**	**Rec.**	**Lic.**
pix2pixHD (T.-C. Wang, Liu, Zhu, Tao, et al., 2018)	High-resolution conditional GAN (semantic-to-image)	✓	✓	✓	✓	✓
SPADE (Park et al., 2019)	Semantic synthesis with SPADE (label/segmentation to image)	✓	✓	✓	✓	✓
gans-traffic (Alam, Martens, & Bazilinskyy, 2025)	Recycle-GAN traffic video translation (N2D, D2N)	✓	✓	✓	✓	✓

Table 4: Representative datasets used in day and night outdoor vision. "Pairing" denotes correspondences across conditions or modalities. Protocol codes follow Table 1. The evaluation column prioritises task aligned evidence rather than image realism metrics alone.

Dataset (ref.)	**Platform**	**Pairing**	**Labels**	**Sensors**	**Protocol codes**	**Most defensible evaluation**
Dark Zurich (Sakaridis et al., 2019)	Driving (images)	Cross-time correspondences	Semantic masks; uncertainty-aware evaluation	RGB	P S D	Segmentation: mIoU and uncertainty-aware IoU; translation: perceptual similarity on correspondences (where used); downstream robustness on real night.
ACDC (Sakaridis et al., 2025)	Driving (images)	Normal-condition correspondences for adverse images	Panoptic labels; uncertainty mask	RGB	S D	Dense perception: panoptic quality, mIoU, detection mAP; uncertainty-aware segmentation when ambiguous regions are present.
Nighttime Driving (Dai & Gool, 2018)	Driving (images)	Unpaired collection	Coarse semantic masks	RGB	S D	Segmentation: mIoU; adaptation gains measured on real night, not only translated imagery.
NightCity (Tan et al., 2021)	Driving (images)	Unpaired collection	Fine semantic masks	RGB	S D	Segmentation: mIoU; report sensitivity to exposure and illumination modelling.
BDD100K (Yu et al., 2020)	Driving (video)	Unpaired collection	Multi-task labels (for example boxes, lanes, drivable area, tracking)	RGB	S D T	Detection and tracking: mAP and tracking metrics; segmentation: mIoU; stratify results by night-time and adverse conditions.

Continued on next page

Dataset (ref.)	Platform	Pairing	Labels	Sensors	Protocol codes	Most defensible evaluation
nuScenes (Caesar et al., 2020a)	Driving (video)	Unpaired collection	Three-dimensional boxes; tracking	Multi-sensor suite	D T	Detection and tracking: dataset-defined metrics (for example nuScenes detection score and mAP); analyse illumination subsets when used for day and night robustness claims.
Waymo Open Dataset (Sun et al., 2020)	Driving (video)	Unpaired collection	Two-dimensional and three-dimensional boxes; tracking	Multi-sensor suite	D T	Detection and tracking: dataset-defined detection and tracking metrics; report condition stratification when evaluating day and night robustness.
Oxford RobotCar (Maddern et al., 2017)	Driving (repeated traversals)	Repeated route across conditions	Ground-truth poses and odometry (with derived tasks in follow-up work)	Multi-sensor suite	S D T	Localisation: pose error and recall under changing conditions; place recognition performance across day and night traversals.
RobotCar Seasons (Sattler et al., 2018)	Driving (localisation benchmark)	Query images registered to a prior map	Six degree-of-freedom camera poses	RGB	S D	Localisation: recall within standard error thresholds and median pose error; report results separately for day and night queries.
Aachen Day-Night (Sattler et al., 2018)	Handheld or mobile	Query images registered to a prior map	Six degree-of-freedom camera poses	RGB	S D	Localisation: recall and pose error for day and night queries; matchability-oriented analysis when translation is used.

Continued on next page

Dataset (ref.)	**Platform**	**Pairing**	**Labels**	**Sensors**	**Protocol codes**	**Most defensible evaluation**
NightOwls (Neumann et al., 2018)	Static surveillance (video)	Unpaired collection	Pedestrian boxes; tracking identifiers	RGB	D T	Detection and tracking: pedestrian mAP and tracking stability; sensitivity to small and partially illuminated pedestrians.
ExDark (Loh & Chan, 2019)	Mixed scenes (images)	Unpaired collection	Object boxes; image-level labels	RGB	D	Detection: mAP; condition-wise analysis across low-light regimes (avoid aggregated reporting only).
KAIST Multispectral Pedestrians (Hwang et al., 2015)	Driving (video)	Paired colour and thermal pairs	Pedestrian boxes; temporal correspondence	RGB and thermal	P D T	Detection: mAP for colour-only, thermal-only, and fused models; report day and night splits explicitly.
LLVIP (X. Jia et al., 2021)	Low-light scenes (images)	Paired visible and infrared pairs	Pedestrian boxes	Visible and infrared	P D	Detection: mAP with explicit alignment assumptions; fusion and translation ablations with consistent downstream evaluation.
FLIR ADAS (Teledyne FLIR LLC, 2018)	Driving (video)	Time-synchronised thermal and visible frames	Object boxes (thermal-oriented annotations)	Thermal and visible	P D T	Detection: mAP; cross-modal robustness reporting; day and night condition analysis where available.

Continued on next page

Dataset (ref.)	**Platform**	**Pairing**	**Labels**	**Sensors**	**Protocol codes**	**Most defensible evaluation**
D^2-City (Che et al., 2019)	Driving (video)	Unpaired collection (condition diversity incl. night)	2D boxes; tracking identifiers (subset fully tracked)	RGB	D T	Detection and tracking: mAP and MOT metrics (for example IDF1/MOTA); stratify by night and adverse conditions; report long-tail cases (glare, blur, heavy traffic).
nuImages (Caesar et al., 2020b)	Driving (images)	Unpaired collection (includes night)	2D boxes; instance masks; semantic segmentation masks	RGB	S D	2D perception: box AP and mask AP; semantic segmentation mIoU; report condition stratification (day/night, rain/snow) when claiming illumination robustness.
Tokyo 24/7 (Torii et al., 2015)	Handheld / street-view retrieval	Same-place queries across day / sunset / night	Place IDs; retrieval ground truth	RGB	S D	Place recognition: Recall@K (and mAP where used); report D2N and N2D separately; analyse failure under extreme lighting and viewpoint mismatch.
Gardens Point Walking (Sünderhauf et al., 2015)	Handheld walking traversals	Sequence correspondence (day vs night; viewpoint change)	Sequence-level correspondences (retrieval pairing)	RGB (night often contrast-enhanced)	S D T	Place recognition: Recall@K under day/night and viewpoint shifts; report robustness under exposure/-contrast preprocessing assumptions.

Continued on next page

Dataset (ref.)	Platform	Pairing	Labels	Sensors	Protocol codes	Most defensible evaluation
MSLS (Mapillary Street-Level Sequences) (Warburg et al., 2020)	Street-level place recognition (large-scale)	Sequence-based clustering; includes day/night test cases	Geo/cluster labels; retrieval ground truth	RGB	S D T	Place recognition: Recall@K with standard MSLS thresholds; report performance on D2N split separately from overall score; include ablations on sequence aggregation.
CVC-14 (González et al., 2016)	Driving / roadside (video)	Visible–FIR paired streams (approx. synchronised)	Pedestrian boxes	Visible and FIR	P D T	Pedestrian detection: mAP for visible-only, FIR-only, and fusion; report day and night splits explicitly and quantify cross-modal alignment sensitivity.
UAVDark135 (B. Li et al., 2022)	UAV tracking (night)	Unpaired collection (night-only benchmark)	Tracking boxes (dense across frames)	RGB	D T	Tracking: success/precision (AUC), robustness to motion blur and low illumination; report per-attribute breakdown and temporal stability.

References

Alam, M.S., Martens, M.H., Bazilinskyy, P. (2025). Generating realistic traffic scenarios: A deep learning approach using generative adversarial networks (gans). *13th international conference on human interaction & emerging technologies: Artificial intelligence & future applications, ihiet-ai 2025* (pp. 349–358). Retrieved from https://doi.org/10.54941/ahfe1005927

Alam, M.S., Parmar, S.H., Martens, M.H., Bazilinskyy, P. (2025). Deep learning approach for realistic traffic video changes across lighting and weather conditions. *2025 8th international conference on information and computer technologies (icict)* (p. 180-185). Hilo, HI, USA. Retrieved from https://doi.org/10.1109/ICICT64582.2025.00034

Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., Van Gool, L. (2019). Night-to-day image translation for retrieval-based localization. *2019 international conference on robotics and automation (icra)* (pp. 5958–5964). Retrieved from https://doi.org/10.1109/ICRA.2019.8794387

Anwar, S., Tahir, M., Li, C., Mian, A., Khan, F.S., Muzaffar, A.W. (2025). Image colorization: A survey and dataset. *Information Fusion*, *114*, 102720, Retrieved from https://doi.org/10.1016/j.inffus.2024.102720

Bang, G., Lee, J., Endo, Y., Nishimori, T., Nakao, K., Kamijo, S. (2024). Semantic and geometric-aware day-to-night image translation network. *Sensors*, *24*(4), 1339, Retrieved from https://doi.org/10.3390/s24041339

Barron, J.T., & Malik, J. (2012). Color constancy, intrinsic images, and shape estimation. *European conference on computer vision* (pp. 57–70). Retrieved from https://doi.org/10.1007/978-3-642-33765-9_5

Barrow, H., Tenenbaum, J., Hanson, A., Riseman, E. (1978). Recovering intrinsic scene characteristics. *Comput. vis. syst*, *2*(3-26), 2, Retrieved from https://doi.org/10.5555/2968618.2968788

Bhattacharjee, D., Kim, S., Vizier, G., Salzmann, M. (2020). Dunit: Detection-based unsupervised image-to-image translation. *2020 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 4786-4795). Retrieved from https://doi.org/10.1109/CVPR42600.2020.00484

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, , Retrieved from https://doi.org/10.48550/arXiv.1801.01401

Blau, Y., & Michaeli, T. (2018). The perception-distortion tradeoff. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6228–6237). Retrieved from https://doi.org/10.48550/arXiv.1711.06077

Buchsbaum, G. (1980). A spatial processor model for object colour perception. *Journal of the Franklin institute*, *310*(1), 1–26, Retrieved from https://doi.org/10.1016/0016-0032(80)90058-7

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., … Beijbom, O. (2020a). nuscenes: A multimodal dataset for autonomous driving. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11621–11631). Retrieved from https://doi.org/10.1109/CVPR42600.2020.01164

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., … Beijbom, O. (2020b). nuscenes: A multimodal dataset for autonomous driving. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11621–11631). Retrieved from https://doi.org/10.1109/CVPR42600.2020.01164

Che, Z., Li, G., Li, T., Jiang, B., Shi, X., Zhang, X., … Ye, J. (2019). D ̉2-city: a large-scale dashcam video dataset of diverse traffic scenarios. *arXiv preprint arXiv:1904.01975*, , Retrieved from https://doi.org/10.48550/arXiv.1904.01975

Chen, Y., Pan, Y., Yao, T., Tian, X., Mei, T. (2019). Mocycle-gan: Unpaired video-to-video translation. *Proceedings of the 27th acm international conference on multimedia* (p. 647–655). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3343031.3350937

Cherian, A., & Sullivan, A. (2019). Sem-gan: Semantically-consistent image-to-image translation. *2019 ieee winter conference on applications of computer vision (wacv)* (pp. 1797–1806). Retrieved from https://doi.org/10.1109/WACV.2019.00196

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., … Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3213–3223). Retrieved from https://doi.org/10.48550/arXiv.1604.01685

Dai, D., & Gool, L.V. (2018). Dark model adaptation: Semantic image segmentation from daytime to nighttime. *2018 21st international conference on intelligent transportation systems (itsc)* (p. 3819–3824). IEEE Press. Retrieved from https://doi.org/10.1109/ITSC.2018.8569387

Du, H., Shi, H., Zeng, D., Zhang, X.-P., Mei, T. (2022). The elements of end-to-end deep face recognition: A survey of recent advances. *ACM computing surveys (CSUR)*, *54*(10s), 1–42, Retrieved from https://doi.org/10.1145/3507902

Durand, F., & Dorsey, J. (2002). Fast bilateral filtering for the display of high-dynamic-range images. *Proceedings of the 29th annual conference on computer graphics and interactive techniques* (pp. 257–266). Retrieved from https://doi.org/10.1145/566654.566574

Ebel, P., Bazilinskyy, P., Colley, M., Goodridge, C.M., Hock, P., Janssen, C.P., … Wintersberger, P. (2024). Changing lanes toward open science: Openness and transparency in automotive user research. *Proceedings of the 16th international conference on automotive user interfaces and interactive vehicular applications* (p. 94–105). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3640792.3675730

Elad, M. (2005). Retinex by two bilateral filters. *International conference on scale-space theories in computer vision* (pp. 217–229). Retrieved from https://doi.org/10.1007/11408031_19

Elad, M., Kimmel, R., Shaked, D., Keshet, R. (2003). Reduced complexity retinex algorithm via the variational approach. *Journal of visual communication and image representation*, *14*(4), 369–388, Retrieved from https://doi.org/10.1016/S1047-3203(03)00045-2

Finlayson, G.D., & Trezzi, E. (2004). Shades of gray and colour constancy. *Color and imaging conference* (Vol. 12, pp. 37–41). Retrieved from https://doi.org/10.2352/CIC.2004.12.1.art00008

Forsyth, D.A. (1990). A novel algorithm for color constancy. *International Journal of Computer Vision*, *5*(1), 5–35, Retrieved from https://doi.org/10.1007/BF00056770

Gatys, L.A., Ecker, A.S., Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, , Retrieved from https://doi.org/10.48550/arXiv.1508.06576

Gehler, P.V., Rother, C., Blake, A., Minka, T., Sharp, T. (2008). Bayesian color constancy revisited. *2008 ieee conference on computer vision and pattern recognition* (pp. 1–8). Retrieved from https://doi.org/10.1109/CVPR.2008.4587765

González, A., Fang, Z., Socarras, Y., Serrat, J., Vázquez, D., Xu, J., López, A.M. (2016). Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, *16*(6), 820, Retrieved from https://doi.org/10.3390/s16060820

Gundersen, O.E., Coakley, K., Kirkpatrick, C., Gil, Y. (2022). Sources of irreproducibility in machine learning: A review.. Retrieved from https://arxiv.org/abs/2204.07610

Gundersen, O.E., & Kjensmo, S. (2018). State of the art: reproducibility in artificial intelligence. *Proceedings of the thirty-second aaai conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth aaai symposium on educational advances in artificial intelligence.* AAAI Press. Retrieved from https://doi.org/10.5555/3504035.3504236

Guo, J., Li, J., Fu, H., Gong, M., Zhang, K., Tao, D. (2022). Alleviating semantics distortion in unsupervised low-level image-to-image translation via structure consistency constraint. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 18249–18259). Retrieved from https://doi.org/10.1109/CVPR52688.2022.01771

Guo, J., Ma, J., García-Fernández, Á.F., Zhang, Y., Liang, H. (2023). A survey on image enhancement for low-light images. *Heliyon*, *9*(4), , Retrieved from https://doi.org/10.1016/j.heliyon.2023.e14558

Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control (MAQC) Society Board of Directors, Waldron, L., … Aerts, H.J.W.L. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), E14–E16, Retrieved from https://doi.org/10.1038/s41586-020-2766-y

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Proceedings of the 31st international conference on neural information processing systems* (p. 6629–6640). Red Hook, NY, USA: Curran Associates Inc. Retrieved from https://doi.org/10.5555/3295222.3295408

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., … Darrell, T. (2018, 10–15 Jul). CyCADA: Cycle-consistent adversarial domain adaptation. J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 1989–1998). PMLR. Retrieved from https://doi.org/10.48550/arXiv.1711.03213

Hogervorst, M.A., & Toet, A. (2008). Method for applying daytime colors to nighttime imagery in realtime. *Multisensor, multisource information fusion: Architectures, algorithms, and applications 2008* (Vol. 6974, pp. 25–33). Retrieved from https://doi.org/10.1117/12.776648

Hoyez, H., Schockaert, C., Rambach, J., Mirbach, B., Stricker, D. (2022). Unsupervised image-to-image translation: A review. *Sensors*, *22*(21), 8540, Retrieved from https://doi.org/10.3390/s22218540

Huang, F., Yao, Z., Zhou, W. (2023). Dtbs: Dual-teacher bi-directional self-training for domain adaptation in nighttime semantic segmentation. *Frontiers in artificial intelligence and applications.* IOS Press. Retrieved from https://doi.org/10.3233/FAIA230382

Huang, S., Jin, X., Jiang, Q., Liu, L. (2022). Deep learning for image colorization: Current and future prospects. *Engineering Applications of Artificial Intelligence*, *114*, 105006, Retrieved from https://doi.org/10.1016/j.engappai.2022.105006

Huang, X., Liu, M.-Y., Belongie, S., Kautz, J. (2018). Multimodal unsupervised image-to-image translation. *Proceedings of the european conference on computer vision (eccv)* (pp. 172–189). Retrieved from https://doi.org/10.1007/978-3-030-01219-9_11

Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1037–1045). Retrieved from https://doi.org/10.1109/CVPR.2015.7298706

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1125–1134). Retrieved from https://doi.org/10.1109/CVPR.2017.632

Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W. (2021). Llvip: A visible-infrared paired dataset for low-light vision. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 3496–3504). Retrieved from https://doi.org/10.1109/ICCVW54120.2021.00389

Jia, Z., Yuan, B., Wang, K., Wu, H., Clifford, D., Yuan, Z., Su, H. (2021). Semantically robust unpaired image translation for data with unmatched semantics statistics. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 14273–14283). Retrieved from https://doi.org/10.1109/ICCV48922.2021.01401

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., ... Wang, Z. (2021). Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, *30*, 2340–2349, Retrieved from https://doi.org/10.1109/TIP.2021.3051462

Jobson, D., Rahman, Z., Woodell, G. (1997, July). A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, *6*(7), 965-976, Retrieved from https://doi.org/10.1109/83.597272

Kennerley, M., Wang, J.-G., Veeravalli, B., Tan, R.T. (2023). 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11484–11493). Retrieved from https://doi.org/10.1109/CVPR52729.2023.01105

Kim, S., Baek, J., Park, J., Kim, G., Kim, S. (2022). Instaformer: Instance-aware image-to-image translation with transformer. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 18321–18331). Retrieved from https://doi.org/10.1109/CVPR52688.2022.01778

Kimmel, R., Elad, M., Shaked, D., Keshet, R., Sobel, I. (2003). A variational framework for retinex. *International Journal of computer vision*, *52*(1), 7–23, Retrieved from https://doi.org/10.1023/A:1022314423998

Kwak, J.-g., Jin, Y., Li, Y., Yoon, D., Kim, D., Ko, H. (2021). Adverse weather image translation with asymmetric and uncertainty-aware gan. *arXiv preprint arXiv:2112.04283*, , Retrieved from https://doi.org/10.48550/arXiv.2112.04283

Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Yumer, E., Yang, M.-H. (2018). Learning blind video temporal consistency. *Proceedings of the european conference on computer vision (eccv)* (pp. 170–185). Retrieved from https://doi.org/10.1007/978-3-030-01267-0_11

Lakmal, H., Dissanayake, M.B., Aramvith, S. (2024). Light the way: an enhanced generative adversarial network framework for night-to-day image translation with improved quality. *IEEE Access*, , Retrieved from https://doi.org/10.1109/ACCESS.2024.3491792

Lan, G., Yang, Y., Wang, Z., Wang, D., Zhao, B., Li, X. (2024). Night-to-day translation via illumination degradation disentanglement. *arXiv preprint arXiv:2411.14504*, , Retrieved from https://doi.org/10.48550/arXiv.2411.14504

Lan, G., Zhao, B., Li, X. (2023). Disentangled contrastive image translation for nighttime surveillance. *arXiv preprint arXiv:2307.05038*, , Retrieved from https://doi.org/10.48550/arXiv.2307.05038

Land, E.H., & McCann, J.J. (1971). Lightness and retinex theory. *Journal of the Optical society of America*, *61*(1), 1–11, Retrieved from https://doi.org/10.1364/JOSA.61.000001

Lee, E., & Kang, S. (2021). Daydriex: Translating nighttime scenes towards daytime driving experience at night. *Applied Sciences*, *11*(5), 2013, Retrieved from https://doi.org/10.3390/app11052013

Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H. (2018). Diverse image-to-image translation via disentangled representations. *Proceedings of the european conference on computer vision (eccv)* (pp. 35–51). Retrieved from https://doi.org/10.1007/s11263-019-01284-z

Lee, Y.-J., Go, Y.-H., Lee, S.-H., Son, D.-M., Lee, S.-H. (2025). Night-to-day image translation with road light attention training for traffic information detection. *Mathematics*, *13*(18), 2998, Retrieved from https://doi.org/10.3390/math13182998

Lengyel, A., Garg, S., Milford, M., van Gemert, J.C. (2021). Zero-shot day-night domain adaptation with a physics prior. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 4399–4409). Retrieved from https://doi.org/10.1109/ICCV48922.2021.00436

Li, B., Fu, C., Ding, F., Ye, J., Lin, F. (2022). All-day object tracking for unmanned aerial vehicle. *IEEE Transactions on Mobile Computing*, *22*(8), 4515–4529, Retrieved from https://doi.org/10.1109/TMC.2022.3162892

Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.-M., Gu, J., Loy, C.C. (2021). Low-light image and video enhancement using deep learning: A survey. *IEEE transactions* *on pattern analysis and machine intelligence*, *44*(12), 9396–9416, Retrieved from https://doi.org/10.1109/TPAMI.2021.3126387

Li, M., Huang, H., Ma, L., Liu, W., Zhang, T., Jiang, Y. (2018). Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. *Proceedings of the european conference on computer vision (eccv)* (pp. 184–199). Retrieved from https://doi.org/10.1007/978-3-030-01240-3_12

Li, X., & Guo, X. (2023, January). Spn2d-gan: Semantic prior based night-to-day image-to-image translation. *Trans. Multi.*, *25*, 7621–7634, Retrieved from https://doi.org/10.1109/TMM.2022.3224329

Li, X., Guo, X., Zhang, J. (2022). N2d-gan: A night-to-day image-to-image translator. *2022 ieee international conference on multimedia and expo (icme)* (pp. 1–6). Retrieved from https://doi.org/10.1109/ICME52920.2022.9859906

Liu, H., Cheng, H., Ye, L. (2024). Dnit: enhancing day-night image-to-image translation through fine-grained feature handling (student abstract). *Proceedings of the aaai conference on artificial intelligence* (Vol. 38, pp. 23563–23564). Retrieved from https://doi.org/10.1609/aaai.v38i21.30474

Liu, J., Xu, D., Yang, W., Fan, M., Huang, H. (2021). Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, *129*(4), 1153–1184, Retrieved from https://doi.org/10.1007/s11263-020-01418-8

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International journal of computer vision*, *128*(2), 261–318, Retrieved from https://doi.org/10.1007/s11263-019-01247-4

Liu, M.-Y., Breuel, T., Kautz, J. (2017). Unsupervised image-to-image translation networks. *Proceedings of the 31st international conference on neural information processing systems* (p. 700–708). Red Hook, NY, USA: Curran Associates Inc. Retrieved from https://doi.org/10.5555/3294771.3294838

Liu, S., John, V., Blasch, E., Liu, Z., Huang, Y. (2018). Ir2vi: Enhanced night environmental perception by unsupervised thermal image translation. *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 1153–1160). Retrieved from https://doi.org/10.1109/CVPRW.2018.00160

Liu, S., Long, W., He, L., Li, Y., Ding, W. (2021). Retinex-based fast algorithm for low-light image enhancement. *Entropy*, *23*(6), 746, Retrieved from https://doi.org/10.3390/e23060746

Loh, Y.P., & Chan, C.S. (2019). Getting to know low-light images with the exclusively dark dataset. *Computer vision and image understanding*, *178*, 30–42, Retrieved from https://doi.org/10.1016/j.cviu.2018.10.010

Ma, X., Ouyang, W., Simonelli, A., Ricci, E. (2024, May). 3d object detection from images for autonomous driving: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, *46*(5), 3537–3556, Retrieved from https://doi.org/10.1109/TPAMI.2023.3346386

Maddern, W., Pascoe, G., Linegar, C., Newman, P. (2017). 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, *36*(1), 3–15, Retrieved from https://doi.org/10.1177/0278364916679498

Mallya, A., Wang, T.-C., Sapra, K., Liu, M.-Y. (2020). World-consistent video-to-video synthesis. *Computer vision – eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part viii* (p. 359–378). Berlin, Heidelberg: Springer-Verlag. Retrieved from https://doi.org/10.1007/978-3-030-58598-3_22

Mao, Q., Tseng, H.-Y., Lee, H.-Y., Huang, J.-B., Ma, S., Yang, M.-H. (2022). Continuous and diverse image-to-image translation via signed attribute vectors. *International Journal of Computer Vision*, *130*(2), 517–549, Retrieved from https://doi.org/10.1007/s11263-021-01557-6

Matsushita, Y., Nishino, K., Ikeuchi, K., Sakauchi, M. (2004). Illumination normalization with time-dependent intrinsic images for video surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(10), 1336–1347, Retrieved from https://doi.org/10.1109/CVPR.2003.1211331

Mittal, A., Moorthy, A.K., Bovik, A.C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695-4708, Retrieved from https://doi.org/10.1109/TIP.2012.2214050

Mittal, A., Soundararajan, R., Bovik, A.C. (2013). Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, *20*(3), 209-212, Retrieved from https://doi.org/10.1109/LSP.2012.2227726

Mo, S., Cho, M., Shin, J. (2018). Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, , Retrieved from https://doi.org/10.48550/arXiv.1812.10889

Mohwald, A., Jenicek, T., Chum, O. (2023). Dark side augmentation: Generating diverse night examples for metric learning. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 11153–11163). Retrieved from https://doi.org/10.1109/ICCV51070.2023.01024

Neumann, L., Karg, M., Zhang, S., Scharfenberger, C., Piegert, E., Mistr, S., . . . others (2018). Nightowls: A pedestrians at night dataset. *Asian conference on computer vision* (pp. 691–705). Retrieved from https://doi.org/10.1007/978-3-030-20887-5_43

Ning, J., & Gong, M. (2023). Enhancing night-to-day image translation with semantic prior and reference image guidance. *Australasian database conference* (pp. 164–182). Retrieved from https://doi.org/10.1007/978-3-031-47843-7_12

Pang, Y., Lin, J., Qin, T., Chen, Z. (2021). Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, *24*, 3859–3881, Retrieved from https://doi.org/10.1109/TMM.2021.3109419

Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. *Computer vision – eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part ix* (p. 319–345). Berlin, Heidelberg: Springer-Verlag. Retrieved from https://doi.org/10.1007/978-3-030-58545-7_19

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. *Proceedings of the ieee/cvf conference* *on computer vision and pattern recognition* (pp. 2337–2346). Retrieved from https://doi.org/10.1109/CVPR.2019.00244

Parmar, G., Park, T., Narasimhan, S., Zhu, J.-Y. (2024). One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, , Retrieved from https://doi.org/10.48550/arXiv.2403.12036

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., … Larochelle, H. (2021, January). Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *J. Mach. Learn. Res.*, *22*(1), , Retrieved from https://doi.org/10.5555/3546258.3546422

Pitié, F. (2020). Advances in colour transfer. *IET Computer Vision*, *14*(6), 304–322, Retrieved from https://doi.org/10.1049/iet-cvi.2019.0920

Pizzati, F., Cerri, P., De Charette, R. (2021). Comogan: continuous model-guided image-to-image translation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 14288–14298). Retrieved from https://doi.org/10.48550/arXiv.2103.06879

Pizzati, F., Lalonde, J.-F., de Charette, R. (2022). Manifest: Manifold deformation for few-shot image translation. *European conference on computer vision* (pp. 440–456). Retrieved from https://doi.org/10.1007/978-3-031-19790-1_27

Punnappurath, A., Abuolaim, A., Abdelhamed, A., Levinshtein, A., Brown, M.S. (2022). Day-to-night image synthesis for training nighttime neural isps. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10769–10778). Retrieved from https://doi.org/10.1109/CVPR52688.2022.01050

Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P. (2002). Color transfer between images. *IEEE Computer graphics and applications*, *21*(5), 34–41, Retrieved from https://doi.org/10.1109/38.946629

Rivoir, D., Pfeiffer, M., Docea, R., Kolbinger, F., Riediger, C., Weitz, J., Speidel, S. (2021). Long-term temporally consistent unpaired video translation from simulated surgical 3d data. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 3343–3353). Retrieved from https://doi.org/10.1109/ICCV48922.2021.00333

Ruder, M., Dosovitskiy, A., Brox, T. (2016). Artistic style transfer for videos. *German conference on pattern recognition* (pp. 26–36). Retrieved from https://doi.org/10.1007/978-3-319-45886-1_3

Sakaridis, C., Dai, D., Gool, L.V. (2019). Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 7374–7383). Retrieved from https://doi.org/10.1109/ICCV.2019.00747

Sakaridis, C., Wang, H., Li, K., Zurbr¨ugg, R., Jadon, A., Abbeloos, W., … Dai, D. (2025). Acdc: the adverse conditions dataset with correspondences for robust semantic driving scene perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-18, Retrieved from https://doi.org/10.1109/TPAMI.2025.3633063

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., … others (2018). Benchmarking 6dof outdoor visual localization in changing conditions. *Proceedings of the ieee conference* *on computer vision and pattern recognition* (pp. 8601–8610). Retrieved from https://doi.org/10.1109/CVPR.2018.00897

Saxena, S., & Teli, M.N. (2021). Comparison and analysis of image-to-image generative adversarial networks: a survey. *arXiv preprint arXiv:2112.12625*, , Retrieved from https://doi.org/10.48550/arXiv.2112.12625

Schutera, M., Hussein, M., Abhau, J., Mikut, R., Reischl, M. (2020). Night-to-day: Online image-to-image translation for object detection within autonomous driving by night. *IEEE Transactions on Intelligent Vehicles*, *6*(3), 480–489, Retrieved from https://doi.org/10.1109/TIV.2020.3039456

Shanaka, I. (2025). *N2D250K: Pixel-to-pixel paired night–day image dataset.* GitHub repository. Retrieved from https://github.com/isurushanaka/N2D250K (Dataset of 250,000 paired night–day image pairs. Accessed 2025-12-19.)

Shiotsuka, D., Lee, J., Endo, Y., Javanmardi, E., Takahashi, K., Nakao, K., Kamijo, S. (2022). Gan-based semantic-aware translation for day-to-night images. *2022 ieee international conference on consumer electronics (icce)* (pp. 1–6). Retrieved from https://doi.org/10.1109/ICCE53296.2022.9730532

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., … others (2020). Scalability in perception for autonomous driving: Waymo open dataset. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2446–2454). Retrieved from https://doi.org/10.1109/CVPR42600.2020.00252

Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., Milford, M. (2015). Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Robotics: Science and Systems XI*, 1–10, Retrieved from https://doi.org/10.15607/RSS.2015.XI.022

Tan, X., Xu, K., Cao, Y., Zhang, Y., Ma, L., Lau, R.W.H. (2021, January). Night-time scene parsing with a large real dataset. *Trans. Img. Proc.*, *30*, 9085–9098, Retrieved from https://doi.org/10.1109/TIP.2021.3122004

Taufiq, M.F.R., & Rahadianti, L. (2025). Cyclegan for day-to-night image translation: a comparative study. *IAES International Journal of Artificial Intelligence*, *14*(3), 2347–2357, Retrieved from https://doi.org/10.11591/ijai.v14.i3.pp2347-2357

Teledyne FLIR LLC (2018). *Teledyne FLIR thermal dataset for algorithm training (flir adas dataset).* Dataset. Retrieved from https://oem.flir.com/solutions/automotive/adas-dataset-form/ (Accessed 18 December 2025)

Tian, Y., Pei, K., Jana, S., Ray, B. (2018). Deeptest: automated testing of deep-neural-network-driven autonomous cars. *Proceedings of the 40th international conference on software engineering* (p. 303–314). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3180155.3180220

Toet, A. (2003). Color the night: applying daytime colors to nighttime imagery. *Enhanced and synthetic vision 2003* (Vol. 5081, pp. 168–178). Retrieved from https://doi.org/10.1117/12.484800

Toet, A., & Hogervorst, M.A. (2012). Progress in color night vision. *Optical Engineering*, *51*(1), 010901–010901, Retrieved from https://doi.org/10.1117/1.OE.51.1.010901

Torbunov, D., Huang, Y., Yu, H., Huang, J., Yoo, S., Lin, M., … Ren, Y. (2023).

Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 702–712). Retrieved from https://doi.org/10.1109/WACV56688.2023.00077

Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T. (2015). 24/7 place recognition by view synthesis. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1808–1817). Retrieved from https://doi.org/10.1109/CVPR.2015.7298790

Van De Weijer, J., Gevers, T., Gijsenij, A. (2007). Edge-based color constancy. *IEEE Transactions on image processing*, *16*(9), 2207–2214, Retrieved from https://doi.org/10.1109/TIP.2007.901808

Wang, H., Li, S.Z., Wang, Y., Zhang, J. (2004). Self quotient image for face recognition. *2004 international conference on image processing, 2004. icip'04.* (Vol. 2, pp. 1397–1400). Retrieved from https://doi.org/10.1109/ICIP.2004.1419763

Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F. (2022). Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, , Retrieved from https://doi.org/10.48550/arXiv.2205.12952

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B. (2018). Video-to-video synthesis. *Conference on neural information processing systems (neurips).* Retrieved from https://doi.org/10.48550/arXiv.1808.06601

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8798–8807). Retrieved from https://doi.org/10.1109/CVPR.2018.00917

Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., Civera, J. (2020). Mapillary street-level sequences: A dataset for lifelong place recognition. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2626–2635). Retrieved from https://doi.org/10.1109/CVPR42600.2020.00270

Wei, X., Zhu, J., Feng, S., Su, H. (2018). Video-to-video translation with global temporal consistency. *Proceedings of the 26th acm international conference on multimedia* (p. 18–25). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3240508.3240708

Wu, X., Wu, Z., Guo, H., Ju, L., Wang, S. (2021). Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 15769–15778). Retrieved from https://doi.org/10.1109/CVPR46437.2021.01551

Xu, Q., Ma, Y., Wu, J., Long, C., Huang, X. (2021). Cdada: A curriculum domain adaptation for nighttime semantic segmentation. *2021 ieee/cvf international conference on computer vision workshops (iccvw)* (p. 2962-2971). Retrieved from https://doi.org/10.1109/ICCVW54120.2021.00331

Yang, S., Sun, M., Lou, X., Yang, H., Zhou, H. (2023). An unpaired thermal infrared image translation method using gma-cyclegan. *Remote Sensing*, *15*(3), 663, Retrieved from https://doi.org/10.3390/rs15030663

Yang, S., Zhou, Y., Liu, Z., Loy, C.C. (2024). Fresco: Spatial-temporal correspondence for zero-shot video translation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8703–8712). Retrieved from https://doi.org/10.1109/CVPR52733.2024.00831

Yi, Z., Zhang, H., Tan, P., Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. *Proceedings of the ieee international conference on computer vision* (pp. 2849–2857). Retrieved from https://doi.org/10.1109/ICCV.2017.310

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., ... Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2636–2645). Retrieved from https://doi.org/10.1109/cvpr42600.2020.00271

Zhang, L., Chen, X., Tu, X., Wan, P., Xu, N., Ma, K. (2022). Wavelet knowledge distillation: Towards efficient image-to-image translation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12464–12474). Retrieved from https://doi.org/10.48550/arXiv.2203.06321

Zhang, M., Zhang, Y., Zhang, L., Liu, C., Khurshid, S. (2018). Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. *2018 33rd ieee/acm international conference on automated software engineering (ase)* (p. 132-142). Retrieved from https://doi.org/10.1145/3238147.3238187

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 586–595). Retrieved from https://doi.org/10.1109/CVPR.2018.00068

Zheng, Y., Dong, W., Blasch, E.P. (2012). Qualitative and quantitative comparisons of multispectral night vision colorization techniques. *Optical Engineering*, *51*(8), 087004–087004, Retrieved from https://doi.org/10.1117/1.OE.51.8.087004

Zheng, Z., Wu, Y., Han, X., Shi, J. (2020). Forkgan: Seeing into the rainy night. *European conference on computer vision* (pp. 155–170). Retrieved from https://doi.org/10.1007/978-3-030-58580-8_10

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the ieee international conference on computer vision* (pp. 2223–2232). Retrieved from https://doi.org/10.1109/ICCV.2017.244

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E. (2017). Toward multimodal image-to-image translation. *Proceedings of the 31st international conference on neural information processing systems* (p. 465–476). Red Hook, NY, USA: Curran Associates Inc. Retrieved from https://doi.org/10.5555/3294771.3294816

Zhuo, L., Wang, G., Li, S., Wu, W., Liu, Z. (2022). Fast-vid2vid: Spatial-temporal compression for video-to-video synthesis. *European conference on computer vision* (pp. 289–305). Retrieved from https://doi.org/10.1007/978-3-031-19784-0_17

Zuiderveld, K.J. (1994). Contrast limited adaptive histogram equalization. *Graphics gems.* Retrieved from https://api.semanticscholar.org/CorpusID:62707267