Machine Vision and Applications (2026) 37:97
https://doi.org/10.1007/s00138-026-01843-8

RESEARCH

A survey of day-night illumination domain translation for outdoor vision: Methods, datasets, and evaluation protocols

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

Received: 13 February 2026 / Revised: 25 April 2026 / Accepted: 14 May 2026

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 develop an IDT-specific 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 23 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, and we synthesise the literature using an evidence-aligned P/S/D/T protocol that highlights recurring failure modes and evaluation gaps.

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 [1–3]. 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 [4, 5]. 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 [5–8]. Adjacent low-light enhancement literature further underscores the broader impact of adverse nighttime imaging, showing that weak illumination and degraded contrast can also affect perception reliability in related vision settings [9–12]. Blur-aware restoration and blur-specific degradation studies are also relevant here because motion blur and mixed blur remain common in night-time outdoor imaging; representative examples include MB-TaylorFormer V2, DBLRNet, and the MC-Blur benchmark [13–15].

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

✉ Md Shadab Alam
m.s.alam@tue.nl

Priyanshu Singh
priyanshu.asn2003@gmail.com

Pavlo Bazilinskyy
p.bazilinskyy@tue.nl

[1] Eindhoven University of Technology, Eindhoven, North Brabant, The Netherlands

[2] Dr. B.C Roy Engineering College, Durgapur, West Bengal, India

Published online: 17 June 2026

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 [16]. In the specific context of night to day (N2D) and day to night (D2N) translation, paired and unpaired learning frameworks [17, 18] 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 [5] motivates the widespread use of weakly supervised and unpaired approaches, as well as outdoor specific translation systems [19], 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 [5]. Localised light sources introduce strong spatial illumination discontinuities [20], while mixed artificial spectra violate standard colour constancy assumptions [21]. 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 [22]. From a probabilistic perspective, this ambiguity manifests itself as perceptual uncertainty and multimodality [23, 24], explaining why translated outputs may appear visually plausible but remain semantically incorrect. These issues motivate the organising lens adopted in this survey, since perceptual improvements do not reliably translate into downstream performance gains [25], and in the video setting explicit temporal coherence constraints are required to prevent flicker and inconsistent structure between frames [26].

Building on the trade off between perceptual quality and fidelity highlighted by Blau et al. [23] and related observations by Zhang et al. [25], 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 [22] and colour transfer formulations [27], 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 [28] and disentanglement orientated translation [29], as well as supervised methods [30], unpaired methods [19], semantic aware objectives for structure preservation [31], multimodal and fusion orientated strategies [32], and temporal formulations for video [26]. Guided by the evaluation concerns raised by Blau et al. [23] and Zhang et al. [25], 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 [33], image colourisation [34, 35], or generic image-to-image translation [36, 37], this article treats IDT as a distinct problem for outdoor vision [38–40]. In this setting, translation errors can alter semantics in ways that are not captured by perceptual metrics alone [41], mapping is often multimodal [42], and video introduces temporal instability that can undermine downstream pipelines [26]. We therefore organise methods around the constraints used to control these failures and complement the review with an artefact availability audit, a dimension that is often underemphasised in previous surveys [33, 43].

This survey makes five contributions. First, we adapt and operationalise a constraint centric taxonomy for outdoor day-night IDT, linking supervision, night-domain gap factors, families of constraints and priors, and recurring failure modes in a way that supports method comparison and evaluation. Second, we provide structured comparative coverage of 30 methods and 23 datasets, together with explicit inclusion criteria that keep the scope clear and reproducible under heterogeneous reporting conditions. 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 such as using PSNR or SSIM without ensuring geometric alignment. Fourth, we provide an evidence-aligned qualitative synthesis (Section 7.3) that shows where perceptual realism measures (P) diverge from downstream utility and stability (D/T). Fifth, we report an artefact availability audit that records what was checked, the audit snapshot date, and a precise definition of availability.

1.1 Key findings and practical guidance

To make the field's current evidence base explicit, we distil cross-cutting findings from the 30 representative methods summarised in Table 12 and the datasets and protocols in Section 7.

- *Unpaired training dominates, so constraints matter more than pixel losses.* Across the 30 methods, 23/30 are primarily unpaired, 5/30 are paired, with 1/30 few-shot and 1/30 hybrid. This skews the literature toward distributional alignment objectives where semantic faithfulness is not guaranteed without additional constraints.
- *Cycle/identity is common, but semantic anchoring is less so.* Cycle/identity constraints appear in 13/30 methods, while semantic/instance constraints appear in 9/30. Physics-informed and correspondence constraints each appear in 5/30 and 5/30, respectively, and explicit temporal constraints appear in only 3/30.
- *Evaluation evidence is heavily weighted toward realism (P) rather than utility (D) or safety-relevant faithfulness (S/T).* Perceptual/distributional evidence (P) is the primary support in 22/30 papers, but explicit semantic/structural evidence (S) appears in only 7/30. Downstream-task evidence (D) is provided in 9/30, and temporal-stability evidence (T) is reported in only 2/30, even though video deployment is a common motivation.
- *Illumination is always addressed; other night-specific gap factors are under-covered.* All 30 entries target illumination shift (I). Far fewer explicitly target glare (G: 2/30), noise (N: 3/30), adverse weather (W: 4/30), or motion/video effects (M: 3/30), which are often the dominant failure drivers in real outdoor operation.
- *Temporal regularisation is often evaluated incompletely.* Only 3 methods introduce explicit temporal constraints in Table 12, and only 2 of those report dedicated temporal stability evaluation. For video-facing claims, per-frame realism is insufficient; Section 7.2 therefore treats temporal evidence as mandatory.
- *Reproducibility is improving, but deployability is still blocked by missing licences and incomplete recipes.* In our artefact audit of 34 methods (Section 3.3), 29 release code and 22 provide pretrained weights, yet only 21 specify licences and 19 provide reproducibility packages. This gap matters because unclear licensing and incomplete run recipes often prevent practical adoption.*Actionable reading guide*. When selecting or reviewing an IDT method, start with (i) the failure mode it targets (Section 4.3), (ii) the constraint family it adds (Table 12), and (iii) the evidence category it genuinely supports under our P/S/D/T protocol (Table 6). This triad helps prevent over-interpreting perceptual metrics as evidence of downstream robustness.

Guideline for hybrid methods. Some methods span more than one family because they combine multiple priors, auxiliary objectives, or supervision signals. In those cases, we use a two-level rule. In the taxonomy crosswalk (Table 12), we code all materially active constraints so the hybrid structure remains visible. In the narrative discussion, however, we assign the paper to the family corresponding to the dominant modelling commitment or the main claim being advanced. For example, ToDayGAN is discussed as task-conditioned because the localisation objective is the main reason for the translation design, even though the model also uses adversarial and semantic machinery; N2D3 is discussed with physics or degradation-aware methods because explicit degradation handling is central to its formulation, even though correspondence cues are also used; and diffusion-backed methods such as PITI or CycleGAN-Turbo remain in the diffusion or foundation-model family when the pretrained generative prior is the main driver of behaviour, even if additional structural or task anchors are added. This rule keeps the discussion readable without erasing hybrid structure, which is retained explicitly in the multi-column coding of Table 12.

From gap types to algorithmic solutions. The five gap factors are intended not only as descriptors of night-time difficulty, but also as a practical guide to which architectural families are likely to help. Table 1 makes this mapping explicit. In broad terms, GAN-style alignment families are strongest when the main problem is global appearance shift and sufficient scene evidence remains visible; diffusion or foundation-model priors are strongest when realism, rare appearance coverage, or broader target-domain support is needed but require extra anchoring to avoid hallucination; and physics-informed or restoration-oriented families are strongest when degradation structure itself is the main obstacle, for example under under-exposure, noise, or motion-related blur. Semantic or task-aware constraints remain important across all three because they provide the main mechanism for turning gap reduction into verifiable preservation and downstream utility.

1.2 Organising lens 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

Table 1 Gap-type to algorithmic-solution mapping across major architectural paradigms

Gap	GAN/alignment families	Diffusion/foundation-model families	Physics/restoration/heuristic families	Main limitation if used alone
I: illumination shift	Strong for coarse domain alignment and global appearance transfer when scene evidence is still visible; often the default for day-night mapping	Strong when broad target-domain realism and richer prior coverage are needed; useful for hard appearance transfer	Useful when the illumination change can be decomposed or exposure effects can be modelled explicitly	Perceptual improvement alone does not guarantee semantic preservation or downstream benefit
G: glare / saturation / mixed spectra	Can suppress coarse appearance mismatch but often struggle with saturated local regions and spectral inconsistencies without semantic anchors	Can inpaint or regularise heavily corrupted regions, but also increase hallucination risk in low-evidence saturated areas	Often best aligned with explicit glare, bloom, or colour-cast modelling and reflective-surface reasoning	Local artefacts may remain hidden by strong global realism scores
N: sensor noise	Can learn denoising implicitly, but usually only when noise patterns are represented well in training data	Can improve perceptual plausibility under noisy conditions, though computational cost is high	Natural match when degradation structure is noise-dominated or coupled to exposure statistics	Weak evidence for object retention if noise handling is evaluated only visually
W: weather interactions	Can adapt broad appearance statistics across fog, rain, or night-weather combinations, especially in unpaired settings	Can provide richer generative support for rare weather-night combinations, but need strong control	Often useful when degradation components can be separated or approximated physically	Factor entanglement makes cross-paper claims fragile without stratified evaluation
M: motion / blur / video effects	Frame-wise GANs are typically weak unless explicit temporal or correspondence constraints are added	Can stabilise local rendering, but are expensive and still need temporal anchoring for streaming use	Often the most natural family when blur or degradation structure is motion-related; video restoration cues can help	Temporal drift, latency, and non-causal assumptions remain underreported without `T` and runtime evidence

The table is intentionally family-level: it indicates which paradigms are usually well matched to each night-specific gap factor and what limitations remain if that factor is handled without additional anchors or evidence

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):

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

Here $\mathcal{L}_{\text{adv}}$ provides distributional alignment, typically via an adversarial objective [17]; $\mathcal{L}_{\text{cyc}}$ enforces the consistency of the cycle or reconstruction [18]; $\mathcal{L}_{\text{id}}$ promotes the preservation of identity [18]; $\mathcal{L}_{\text{sem}}$ encodes semantic or task-consistency constraints through labels or task networks [44]; $\mathcal{L}_{\text{phy}}$ captures physics-informed priors (e.g. illumination and image-formation structure) [22, 28]; and $\mathcal{L}_{\text{temp}}$ enforces temporal coherence for video translation [26]. 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 [17, 18, 26]. As an organising lens, this formulation is comparative rather than theoretical: 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. It also makes explicit which constraints are used to control night specific ambiguity and provides a shared basis for analysing method families and their evidence requirements.

1.2.1 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 [17].

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

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

where $\mathcal{L}_{\text{adv}}$ denotes the adversarial alignment between translated outputs and the target-domain distribution [17]. 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 [5]. 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 [42, 45, 46]. This can yield outputs that appear visually convincing while deviating from the underlying scene semantics, consistent with the perception distortion trade-off [23]. The risk is amplified by assumptions that are routinely violated at night, including smooth illumination in the presence of localised light sources [20], reduced signal-to-noise ratios under short exposure [4], limited visibility of small or distant semantic cues [8], and Lambertian reflectance and colour constancy under mixed artificial spectra [21]. 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, day-night robustness was commonly addressed with classical illumination modelling and colour processing rather than learned cross-domain mappings. The objective was to reduce sensitivity to illumination by improving visibility, stabilising colour, or normalising lighting so that downstream recognition pipelines degrade less under low illumination.

2.1 Illumination and colour normalisation

A first line of work manipulates contrast and dynamic range to make dark regions more informative. Typical operators include local histogram equalisation variants and tone-mapping style decompositions that compress large-scale illumination while preserving fine-scale detail [47, 48]. While effective for improving perceived visibility, such operators can introduce haloing, amplify sensor noise, and alter colour balance, especially in the lowest-exposure regions.

A second line of work targets illumination invariance through illumination/reflectance separation. Retinex-style reasoning and intrinsic-image formulations view the observed image as reflectance modulated by illumination, and aim to normalise or estimate illumination so that the recovered reflectance is more stable across lighting [22, 49]. Practical Retinex variants introduced multiscale processing and colour restoration to improve stability in real imagery [20, 50].

A complementary classical strategy is global appearance normalisation via statistical colour transfer, which aligns simple colour statistics between images to reduce gross chromatic differences while preserving geometry [27]. For video, illumination can be treated as a time-varying signal; early work proposed temporal illumination normalisation and low-dimensional illumination models to reduce lighting fluctuations in fixed-camera sequences [51, 52].

2.2 Limitations and connection to modern translation

Classical illumination and colour normalisation methods provide interpretable priors, but their assumptions are often violated in real nighttime outdoor scenes. Illumination is not spatially smooth (localised lamps and headlights), colour constancy is weak under mixed artificial spectra, and sensor noise and saturation corrupt the signal in low-exposure regions [5, 20, 21]. Moreover, these pipelines are largely agnostic to semantic structure, so the output can improve visibility while distorting or suppressing small objects and fine textures that matter for safety-critical perception.

A useful link to modern IDT is the simplified image-formation view

$$I(x) = R(x) \cdot L(x), \quad (5)$$

where $R(x)$ encodes scene reflectance/content and $L(x)$ encodes illumination. Classical methods regularise $L(x)$ with hand-crafted priors. Learning-based translation can be interpreted as replacing such priors with data-driven objectives (paired reconstruction or unpaired distribution alignment) while adding explicit constraints (cycle/identity,

semantic/instance, correspondence/contrastive, physics-inspired, and, for video, temporal) to control semantic drift and temporal instability under severe illumination ambiguity [17, 18, 26, 31, 53].

3 Scope, terminology and survey protocol

3.1 Scope and terminology

We survey methods whose primary objective is IDT for outdoor RGB images and RGB video, covering both D2N and N2D settings in support of downstream tasks such as detection, segmentation, localisation, and tracking. Here, translation means mapping an input to a different illumination domain while preserving scene geometry and semantics; for video, temporal coherence is also required.

We distinguish IDT from enhancement and domain adaptation because their validation targets differ. Enhancement improves visibility within the same illumination domain, whereas domain adaptation seeks task robustness across domains without requiring a translated image or video as the final output. Accordingly, we treat enhancement and adaptation as context unless they directly support IDT or contextualise evaluation practice. Cross-modal settings such as thermal-to-visible or infrared-to-visible translation are likewise outside the main method set: they are important for night perception, but they introduce modality shift in addition to illumination shift and are therefore summarised separately in Appendix B. Throughout the survey, IDT claims are judged by whether translated outputs preserve scene content across illumination domains and improve downstream outdoor vision under the `P/S/D/T` protocol.

For transparency, the main exclusions are: (i) pure low-light enhancement methods that do not target day-night translation as a domain-mapping problem, (ii) domain-adaptation methods without explicit translated outputs, (iii) cross-modal thermal or infrared translation settings, and (iv) generic image-to-image baselines that are cited only as references rather than included as IDT methods. These classes are relevant neighbouring literatures, but they are not treated as part of the main comparative set because they differ in objective, validation target, or domain-shift structure.

Scope boundary: illumination shift versus modality shift. Night-time outdoor perception is also studied in *multispectral* settings that involve paired visible–thermal or visible–infrared streams. These benchmarks are valuable for night perception, but they introduce a *sensor modality shift* in addition to illumination change. Because this survey focuses on illumination domain translation for *RGB* imagery and video (day $\leftrightarrow$ night), we treat multispectral translation as *out of scope* and exclude it from the main comparative set.

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.

3.2.1 Authority-weighted curation (Tier A / Tier B)

The day-night translation literature is heterogeneous in venue quality and evidentiary strength. To keep the survey useful as a reference for practitioners, we explicitly separate *authoritative* sources from *contextual* ones. We use this separation in two ways: (i) the strongest comparative statements and the "minimum evidence" decision rules in the main paper are anchored in Tier A sources; and (ii) Tier B sources are included to complete the taxonomy and to document variants, but they are not the sole basis for best-practice recommendations.

- *Tier A (authoritative)*. Peer-reviewed papers in flagship computer vision / machine learning venues and journals

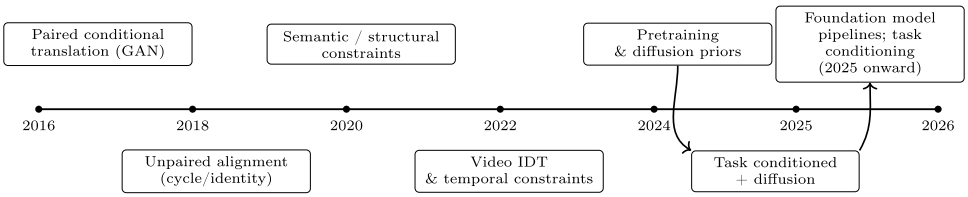

Fig. 1 High level timeline of modelling paradigms for outdoor day-night IDT. The survey uses a curated spine of representative methods in each paradigm to prioritise discussion, to support actionable evaluation guidance, and to highlight emerging integrations such as task conditioned diffusion pipelines

Table 2 Curated "spine" used to prioritise discussion 12

Paradigm	Typical constraints/ priors	Representative anchors	Year	Venue	Minimum evidence
Paired conditional translation	Supervised reconstruction; paired correspondence	Pix2Pix [17], Paired synthetic N2D [30]	2017, 2024	CVPR, IEEE Access	P+S, then D
Unpaired GAN alignment	Cycle/identity; adversarial distribution matching	CycleGAN [18], UVCGAN [54]	2017, 2023	ICCV, WACV	S+D
Semantic preservation	Segmentation/ instance constraints; feature consistency	SPADE-style synthesis [55], SGA-D2N [56]	2019, 2024	CVPR, Sensors	S+D
Task-conditioned translation	Optimise for retrieval and localisation objectives	ToDayGAN [40]	2019	ICRA	D (task)
Physics / degradation disentangling	Illumination decomposition; degradation models	D2N ISP Synthesis [57], N2D3 [58]	2022, 2024	CVPR, arXiv	S+D
Pretraining and diffusion priors	Large-scale pretraining; diffusion priors; hallucination-aware semantic anchoring	PITI [24], Bridging Day and Night [59]	2022, 2026	arXiv, AAAI	S+D under stressors

We use these anchors to differentiate historically important baselines, currently competitive families, and emerging directions, and to ground the decision rules and evidence checklist in the main paper. Full method-level details are provided in Table12

(for example, CVPR/ICCV/ECCV; TPAMI/IJCV; and other long-standing, high-selectivity venues), as well as widely adopted baselines with reproducible artefacts that are regularly used for comparison.

- *Tier B (contextual)*. ArXiv-only preprints, workshop papers, domain-general outlets, and niche variants. These may contain useful ideas, but their claims are treated as provisional unless corroborated by Tier A evidence and/ or independent downstream evaluation.

When a preprint later appears in a peer-reviewed venue, we prioritise and cite the peer-reviewed version where possible. We also separate historically important baselines from currently competitive methods by emphasising a curated "spine" (next subsection) and by reporting evidence codes (P, S, D, T) for each family.

3.2.2 Curated SOTA spine and prioritisation (2016–2026)

To reduce the risk of "flat" coverage where all papers appear equally important, we curate a compact spine of representative methods that capture the dominant modelling paradigms and that serve as anchors for evaluation recommendations. The spine is *not* intended to be exhaustive; instead, it provides a high-signal overview of what is historically foundational, what is currently competitive, and which directions are emerging. Figure 1 provides a high-level timeline of these modelling paradigms. The long tail of variants is still documented in the taxonomy tables, but the spine drives the survey narrative and the decision rules. Table 2 summarises the curated spine used to prioritise the discussion.

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 23 public benchmarks and evaluation resources for outdoor RGB imagery and video that are used in the included IDT papers or are directly relevant to the P/S/D/T protocol as evaluation reference points. To improve timeliness, we updated the search and screening window to include relevant works released up to 2026 and incorporated recent in-scope methods where they provided sufficient methodological detail for comparison. The purpose of this coverage is not exhaustive aggregation for its own sake, but to support a controlled comparative synthesis across constraint families, night-domain gap factors, and evidence types under heterogeneous reporting conditions. In line with our authority-weighted curation, peer-reviewed papers remain

the primary basis for comparative recommendations, while some recent 2026 works, including several preprint-only studies, are included as contextual evidence when they directly inform emerging trends such as diffusion priors, hallucination suppression, and stronger semantic anchoring. Although the procedure is systematic, omissions remain possible due to the breadth of related work, differences in terminology between communities, and the continued pace of new releases.

3.3 Artefact audit protocol

In addition to summarising methods and datasets, we report an artefact availability audit for representative day-night IDT methods covered in this survey. The audit is intended as a transparency aid: it documents what a reader can reasonably expect to find *at the audit snapshot date* (see Table 11) and reduces ambiguity about what "code available" means in practice. This audit is document-based rather than execution-based: it records the publicly available artefacts and metadata needed for reproduction, but it does not claim that every released repository was rerun end-to-end under a standardised verification environment.

Conservative availability rule. An artefact is marked as available (✓) only if it is publicly accessible (e.g. a repository, model card, or archive) and provides enough information for its stated purpose. Broken links, private repositories, partial releases without the required scripts/configuration, or missing licensing information are treated as unavailable.

Table 3 Audit rubric for artefact availability used throughout the availability matrix (Table 10)

Field	Criteria for ✓
Code	Public source code for training/inference (beyond a paper-only description), including key scripts/configs.
Wts.	Downloadable pretrained weights aligned with the described model/version.
Data	Public dataset access and/or released splits/annotations required to evaluate; clear acquisition instructions if restricted.
Rec.	Runnable reproduction package (configs/hparams + environment/deps + commands + evaluation script) sufficient to reproduce the main reported numbers under the stated protocol.
Lic.	Explicit licence terms for code and, when stated, pretrained weights (any licence is acceptable if clearly specified).

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. We do not claim that constraint-aware categorisation is wholly unprecedented in broader GAN, image translation, or domain adaptation literature; rather, our contribution is to specialise this lens for outdoor day-night IDT by coupling it explicitly to night-domain gap factors, typical failure modes, and the evidence requirements captured by our P/S/D/T protocol.

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

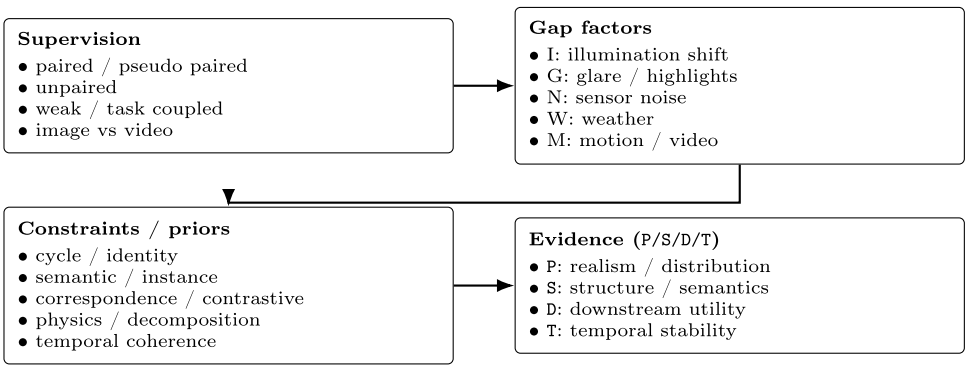

Fig. 2 Constraint centric taxonomy at a glance. We organise IDT work by (i) supervision regime, (ii) the dominant night domain gap factors (I/G/N/W/M), (iii) constraint and prior families introduced to preserve content, and (iv) the evidence provided under the P/S/D/T protocol. The full method level crosswalk is provided in Appendix A

Table 4 Decision-oriented summary by method family

Family	Typical constraints/priors	When it helps/typical failure modes	Minimum evidence
Unpaired GAN baselines	Cyc/Id; adversarial alignment	Improves global appearance; can drift semantically in low evidence regions (small objects, deep shadows) and under glare; requires explicit checks for object retention or hallucination.	P+S; D if claiming benefit
Semantic/ instance anchored	Task network, labels, instance consistency	Improves object integrity and layout preservation; depends on label quality and task network domain robustness; can overfit to a single task model.	S mandatory; D for task claims
Correspondence / contrastive	Patch or feature matching; local consistency	Reduces local distortions; helps texture poor regions; may still fail under severe exposure loss where correspondences are unreliable.	S; D recommended
Physics / decomposition priors	Illumination reflectance structure; degradation modelling	Targets night specific degradations (noise, glare, under exposure); failure under complex non Lambertian effects or mixed lighting if assumptions break.	S+D under stressors
Temporal (video)	Flow or feature coherence; spatio temporal objectives	Reduces flicker or identity drift; fragile when flow fails at night; long horizon drift is often under reported.	T mandatory; D for tracking or localisation
Foundation/ diffusion pipelines	Strong generative priors; often needs anchoring	High realism (P); risk of geometry or identity drift and hallucination without anchors; needs conservative evaluation.	S+D (and T for video)

This table is designed for reading flow; the complete method-level crosswalk is in Table 12

to regularise the mapping under extreme illumination shift. Figure 2 summarises the constraint-centric taxonomy used in this survey.

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 [17, 57]. 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 [18, 60, 61]. 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 [53, 56].

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 [26, 62]. 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 [18].
- *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 [53, 63].
- *Physics-informed priors* encode illumination and degradation structure explicitly-e.g. via decomposition, image formation models, or invariants-reducing reliance on purely statistical alignment [64].
- *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 [29, 65].
- *Temporal constraints* enforce frame-to-frame coherence in video IDT, mitigating flicker, colour instability, and identity drift that degrade tracking and localisation [26].

Table 12 provides a crosswalk from methods to constraints for IDT (provided in Appendix A for readability). 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).

Table 5 Compact dataset shortlist for acceptance-grade evidence. The complete catalog (23 datasets) is in Appendix Table13

Dataset	Best for	Supports	Why it is defensible
Aachen / RobotCar [77, 78]	localisation / retrieval	SD	pose/retrieval ground truth supports matchability- and task-driven evaluation.
ACDC [79]	adverse conditions incl. night	D	multi-condition suite; supports factor-stratified reporting beyond clear-night cases.
BDD100K (night) [7]	multi-task driving video	DT	video labels enable task and temporal evidence; useful for deployment-oriented claims.
CROWD [80]	global urban dashcam video	DT	minute-scale contiguous clips with day/night labels, detector outputs, and segment-local tracks support external robustness and temporal stress testing; not a paired IDT benchmark.
Dark Zurich [39]	driving segmentation under night	SD	cross-time correspondences + uncertainty-aware labels enable structure checks and downstream evaluation.
ExDark [81]	low-light detection	D	enables condition-wise detection reporting under varied low-light regimes.
NightOwls [8]	night pedestrian det./tracking	DT	dedicated night surveillance video with boxes and tracking IDs.

4.2 Literature overview: which constraints and which evidence dominate?

Table 12 makes two field-level patterns explicit.

Constraint mix. Cycle/identity constraints are the most frequent (13/30), while semantic/instance anchoring appears in 9/30. Physics-informed and correspondence constraints each appear in 5/30 and 5/30, and explicit temporal constraints appear in only 3/30. This suggests that many pipelines still rely on generic distribution alignment with weak anchoring to scene content in low-evidence nighttime regions.

Evidence mix. 22/30 papers primarily support claims with perceptual/distributional evidence (P). In contrast, explicit semantic/structural evidence (S) appears in only 7/30, downstream-task evidence (D) in 9/30, and temporal evidence (T) in only 2/30. As a result, a large subset of the literature remains difficult to compare for safety-relevant use-cases where object retention, geometry, and temporal stability are the central requirements. Taken together, these patterns suggest that methodological progress in IDT has outpaced evaluation practice: the field has become better at producing plausible night-domain appearance than at demonstrating robust preservation of semantics, downstream utility, and temporal stability.

This imbalance is not accidental. Semantic and temporal constraints are comparatively scarce in real night scenes because both depend on supervision signals that degrade precisely under the conditions that make IDT difficult. Semantic anchors require reliable labels or pretrained task networks, yet dense nighttime annotations are expensive to obtain and daytime-trained segmentation or detection models often become brittle under low contrast, glare, colour casts, and partial visibility. Temporal constraints face a parallel problem: many formulations assume dependable correspondences across frames, but optical flow, patch matching, and motion cues become unstable at night under noise, motion blur, specular highlights, exposure jumps, and moving light sources. In practice, these two weaknesses compound one another, which helps explain why the literature more often defaults to cycle/identity-style regularisation than to stronger semantic or temporal anchoring in real outdoor night settings.

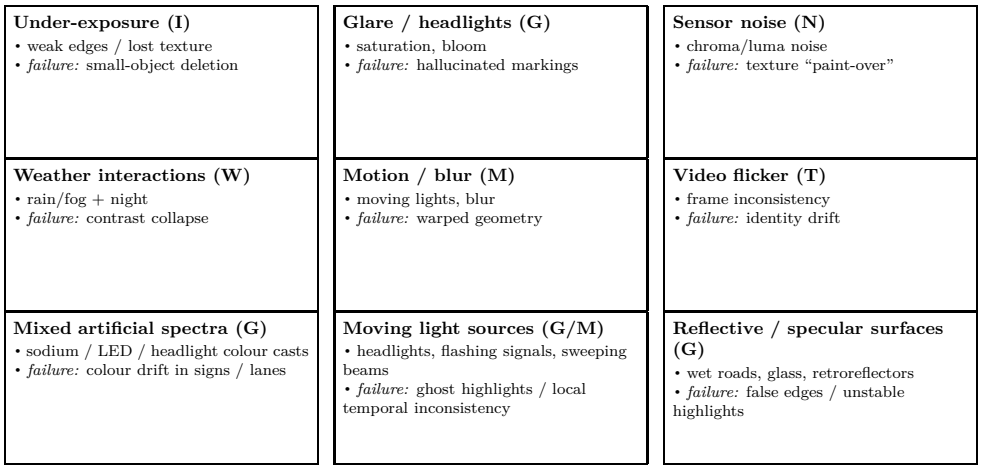

Fig. 3 Visual scaffolding for night-specific domain gaps and common IDT failure modes. Each cell lists (i) the stressor and (ii) a typical translation failure that can mislead downstream outdoor vision. Use this figure as a navigation aid when reading the constraint families and evaluation protocol

4.3 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 [23, 66]. Figure 3 provides a visual scaffold of night-specific domain gaps and common IDT failure modes. 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* [53, 63]. In regions with limited evidence, particularly in N2D translation, generative models can introduce visually plausible but incorrect structure, reflecting ambiguity in conditional mapping and objectives that prioritise perceptual plausibility [23]. Global normalisation and global losses can also suppress weak signals from small or distant objects, motivating local patch or correspondence constraints that preserve structure [29, 65].

Two practically important nighttime stressors warrant more explicit emphasis: mixed artificial spectra and moving light sources. Mixed artificial spectra from sodium lamps, LEDs, traffic signals, and vehicle headlights violate colour-constancy assumptions and can induce spatially varying colour casts, sign or lane misrendering, and instability in reflective surfaces even when global brightness appears plausible. Moving light sources introduce local, time-varying illumination discontinuities that destabilise correspondences and temporal constraints, especially in video, causing ghost highlights, local structure drift, and short-horizon inconsistency. In our taxonomy, these effects are best interpreted as night-specific manifestations within the existing glare and motion categories rather than as separate top-level gap dimensions, but they merit explicit treatment because they are frequent and safety-relevant in outdoor night scenes. Related blur-aware restoration work further reinforces that blur at night is not a single nuisance variable: recent restoration architectures and mixed-cause blur benchmarks show that motion, defocus, and low-light degradation can interact in ways that complicate both translation and evaluation [13–15]. We therefore treat blur-aware restoration as an adjacent reference literature for interpreting motion-related failure modes, even though such methods are not included in the main RGB IDT comparative set. 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 [26, 62].

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.

These failure modes are not merely illustrative: they motivate the minimum evidence requirements developed in Section 7.2, where we turn from failure analysis to operational validation.

4.4 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.

- *Supported*: Under unpaired supervision, distribution alignment plus cycle/identity constraints can improve global appearance (`P`) but does not reliably prevent semantic drift; additional anchoring constraints are typically required for faithfulness (`S`).

- *Promising*: Semantic/instance constraints, correspondence/contrastive objectives, and physics-informed priors reduce failure rates in low-evidence regions and mixed illumination, but results are often not yet validated consistently with downstream evidence (D) across datasets.
- *Unclear*: Which constraint families yield the best worst-case behaviour under glare, severe under-exposure, and sensor-specific noise remains insufficiently settled because most studies under-report S/D/T relative to P.

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 multimodal 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 [17]. 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 [57]. 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 [18]. Closely related formulations include dual learning approaches [60] 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 [61]. 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 [56, 63]. 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 [53]. 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 [29]. 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 [40]. 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 [67]. 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 [58]. 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.

Representative temporal-stability methods and design lines include vid2vid-style flow-guided temporal GANs, motion-aware cycle-based translation, recurrent or three-dimensional video generators, and recent feature-space stabilisation methods such as FRESCO [26, 62, 68, 69]. Although these methods are not all day-night IDT systems in the strict narrow sense, they provide the main technical toolbox currently used to suppress flicker, colour oscillation, and identity drift in video translation and are therefore directly relevant to outdoor night-facing IDT.

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 [54, 70, 71]. Although such refinements can improve visual quality, their impact should be assessed together with semantic preservation and task utility, particularly in safety-relevant applications.

- *Supported*: Generic unpaired translation (adversarial + cycle/identity) often improves perceived brightness/contrast (P) but can remove small objects or distort layout without explicit content anchoring (S).
- *Promising*: Semantic/instance-guided objectives and correspondence/contrastive constraints improve structure preservation in ambiguous night regions; physics-guided and decomposition-style approaches better target mixed illumination, glare, and noise but are less consistently evaluated for downstream gain (D).
- *Unclear*: The conditions under which diffusion/foundation-model pipelines preserve safety-critical semantics without hallucination are not yet well characterised; acceptance-grade claims require explicit S and D evidence under night-specific stressors.

5.8 Diffusion and foundation-model-driven translation

Recent work increasingly leverages large-scale pretraining and diffusion-style priors for image-to-image generation. In the IDT context, these approaches can improve *photorealism* (P) and broaden coverage of rare appearances, but they

also introduce new risks for outdoor systems: hallucinated objects, altered geometry, and inconsistent rendering across frames. Consequently, diffusion or foundation-model pipelines should not be assessed primarily via image realism metrics; they require explicit S and D evidence and stress testing under night-specific failure modes (e.g. glare, low texture, under-exposure, and moving lights).

A practical way to integrate this paradigm into IDT pipelines is to treat pretrained priors as *components* rather than end-to-end guarantees. Diffusion-prior approaches such as PITI [24] can be combined with constraints that anchor structure and content (e.g. correspondence- or semantics-based anchors). More recent work also suggests a shift from realism-oriented generation toward stronger control of semantic failure modes. In particular, recent 2026 work on unpaired day-night translation explicitly targets target-class hallucinations by combining multi-step translation with semantic anchoring mechanisms that detect and suppress hallucinated content in safety-critical categories such as vehicles and traffic lights [59]. This trend reinforces our central claim that strong generative priors are not sufficient by themselves: recent methods are most convincing when they provide evidence of structural faithfulness and downstream utility, rather than relying on perceptual realism alone. In our taxonomy, these methods provide a strong generative prior, while the *constraints* and the *protocol* determine whether the translated outputs are safe to use as pre-processing for downstream tasks.

We also note that some of the newest 2026 contributions are currently available as preprints rather than archival journal or conference versions; accordingly, we treat them as contextual evidence of emerging directions rather than as the sole basis for best-practice recommendations.

- *Supported*: Pretraining and diffusion priors can improve visual plausibility (P) but do not reliably guarantee semantic faithfulness without additional constraints; evaluation must include S and D.
- *Promising*: When combined with explicit structural or semantic anchors, pretrained priors can reduce failure cases in low-texture regions, under extreme exposure, and in hallucination-prone categories.
- *Unclear*: The conditions under which foundation-model pipelines improve D for safety-relevant outdoor tasks remain insufficiently characterised, especially under distributional stressors and long-tail events.

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 [26, 72, 73]. 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 [26, 62]. Recurrent generators and three-dimensional architectures provide an alternative by propagating information over time to reduce frame variance and stabilise repeated structures [26, 68]. 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 [69].

Representative mechanisms can be summarised as follows. In this sense, the survey treats temporal stability not as a marginal implementation detail, but as a distinct methodological axis spanning flow-guided temporal GANs, cycle-based motion-aware translation, recurrent/3D video generators, and representation-level stabilisation methods.

- Optical-flow-guided warping losses that penalise inconsistencies after motion compensation [26, 73].
- Spatio-temporal discriminators and motion-aware objectives that encourage realistic short-term dynamics [26, 62].
- Recurrent and three-dimensional designs that propagate context and reduce frame-to-frame variance [68].
- Feature-space temporal coherence objectives that stabilise intermediate representations [69].

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 [26, 68]. 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 [72, 73]. 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 [62]. Feature-space coherence mitigates some of these issues, but its effectiveness depends on the robustness of the chosen representations under severe illumination variation [69].

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 re-enter the field of view or when lighting changes abruptly. Mechanisms for longer-term consistency have therefore been explored in video synthesis and translation [74, 75], alongside practical acceleration strategies that reduce latency while maintaining temporal stability [76]. 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 [73] and sequence-level stability criteria that quantify drift over time [68]. 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 [26, 68].

- *Supported*. Frame-wise translators applied to video commonly introduce flicker and identity drift; temporal coupling is required for stable deployment (`T`).
- *Promising*. Flow-guided and feature-space temporal regularisation reduce short-term inconsistency, and spatio-temporal discriminators improve global temporal realism, but robustness under night-specific temporal stressors (headlights, exposure jumps) still requires stronger evaluation.
- *Unclear*. Long-horizon stability and the relationship between temporal smoothness and sequence-level downstream performance remain under-studied because `T` and sequence `D` are rarely reported together.

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 the translation objective and what constitutes defensible evaluation: (i) the availability of temporal or cross-time correspondence, (ii) whether annotations exist for downstream tasks, and (iii) whether the data isolate illumination change or mix it with additional factors such as adverse weather or sensor modality shift. For outdoor vision, the most actionable benchmarks are those that provide either explicit cross-time correspondences (to support limited pixel-aligned measures and structure checks) or rich task labels that enable downstream evaluation on real night imagery. Recent dataset releases also suggest a gradual shift toward stronger paired or aligned day-night evaluation in outdoor driving, which may help future studies move beyond weak correspondences and support more defensible structure-preservation and downstream-utility assessment; for example, DarkDriving introduces real-world aligned

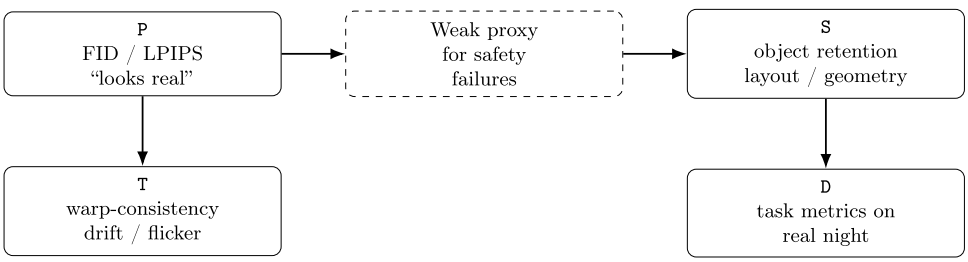

Fig. 4 Evidence mapping: why `P`-only reporting is insufficient for outdoor IDT. Realism/distribution metrics (`P`) are weak proxies for semantic faithfulness and downstream robustness; acceptance-grade claims require matching evidence types (`S`/`D`/`T`) to the claim class

Table 6 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 both FID and KID to avoid bias from a single metric. 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.

Table 6 is designed for IDT validation. Enhancement only or domain adaptation only papers may use overlapping metrics, but they do not satisfy IDT evidence requirements unless they evaluate translated outputs as illumination domain mappings rather than only visibility restoration or task transfer robustness

day-night image pairs for autonomous driving in dark environments [82].

Web-sourced dashcam corpora add a complementary evaluation axis: ecological scale and geographic diversity. CROWD provides 51,753 manually curated, minute-scale urban dashcam segment records from 42,032 YouTube videos, corresponding to 20,275.56 retained hours across 7,103 localities in 238 countries and territories, with segment-level day/night labels, YOLOv11x detections, and BoT-SORT tracks [80]. This makes it useful for external robustness audits, detector/tracker based downstream checks, and temporal stability stress tests, especially when claims extend beyond a single city or sensor pipeline. It should not, however, be treated as a paired day-night translation benchmark: the release indexes third-party videos rather than redistributing them, and the object annotations are machine generated rather than dense human semantic labels.

For dense driving perception, *Dark Zurich* provides cross-time correspondences and supports uncertainty-aware evaluation for semantic segmentation [39]. *ACDC* extends this setting to adverse conditions (including night) with dense labels for perception tasks [79]. *NightCity* and the nighttime subset of *BDD100K* provide unpaired nighttime collections with semantic annotations that enable downstream evidence on real night imagery [7, 83]. For detection under low illumination, datasets such as ExDark provide object-level annotations and are frequently used for condition-wise analyses rather than aggregate reporting alone. For large-scale routine urban driving, *CROWD* is most appropriate as a held-out robustness and temporal-continuity resource, because its day/night labels and derived tracks support stratified downstream checks while its lack of aligned cross-time pairs limits pixel-level translation evaluation [80].

For localisation and retrieval under changing conditions, Aachen Day-Night and RobotCar Seasons provide strong stress tests for viewpoint and illumination changes, and they are frequently used to evaluate whether translation helps (or harms) feature matching and retrieval pipelines [77]. Tokyo 24/7 is similarly used to probe extreme appearance change in place recognition [84]. In video settings, benchmarks with tracking identifiers or sequence-level ground truth enable temporal evidence (T) and downstream robustness (D) in dynamic scenes, which is crucial when translation is proposed as a pre-processing step for tracking or localisation.

7.2 Evaluation protocol and common pitfalls

We now turn from method and failure analysis to the question of what counts as defensible evidence. The central practical issue is *metric mismatch*: commonly reported image metrics are weak predictors of outdoor vision utility. Figure 4 illustrates why P-only reporting is insufficient for outdoor IDT evaluation. PSNR and SSIM are meaningful only under truly aligned pairs, whereas many D2N benchmarks provide approximate, synthetic, or no correspondence at all [39, 79]. When ground truth is unavailable, authors often report perceptual or distributional measures such as FID, KID, and LPIPS [25, 85], but these can still miss local safety failures such as small-object removal, lane deformation, or traffic-sign hallucination. No-reference quality metrics such as NIQE and BRISQUE are likewise hard to interpret because nighttime imagery violates the natural-image statistics on which they were designed [86, 87].

Table 6 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. The purpose of the P/S/D/T protocol is comparative rather than experimental: it provides a common decision framework for interpreting heterogeneous IDT papers according to the strength and relevance of the evidence they report. 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

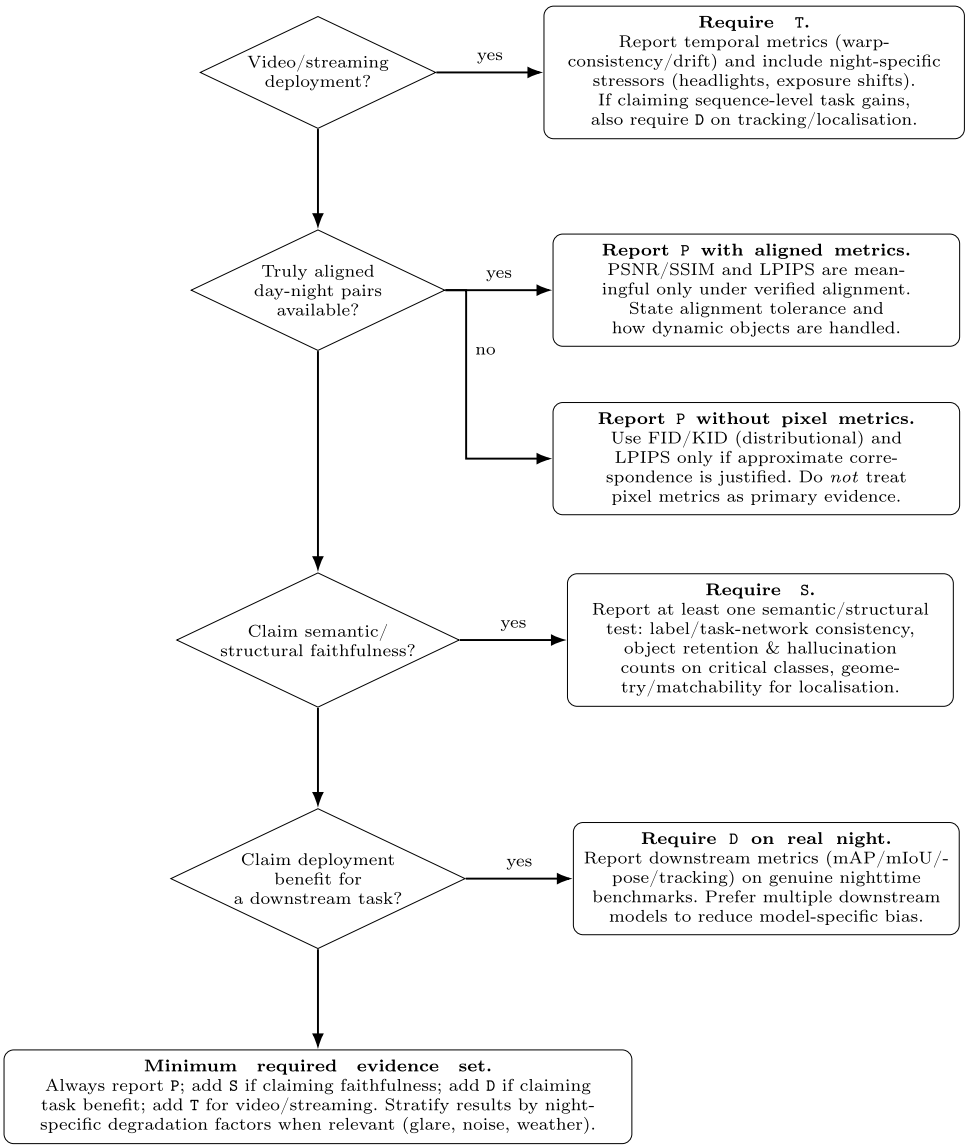

Fig. 5 Protocol selection decision tree for IDT evaluation. The leaves specify which evidence components from Table 6 are *mandatory* given the intended use-case and claim. The goal is to prevent over-claiming from perceptual metrics alone and to make results comparable across papers

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.

7.2.1 Minimum reporting checklist (P/S/D/T)

Protocol selection decision tree. Figure 5 provides a practical decision tree for selecting an *evaluation protocol* (and, implicitly, the minimum evidence components) given the dataset structure and the claims being made. The minimum required evidence set is the union of all components triggered by the applicable branches.

Minimum evidence checklist. Table 7 consolidates a *minimum* evidence checklist into directly actionable requirements. It is designed for both authors and reviewers: each claim class maps to the evidence components that should be treated as mandatory.

Decision rules (falsifiable best practices). To reduce qualitative over-claiming, the following rules can be applied as acceptance-grade reporting standards:

1. *IF* evaluation is unpaired or correspondence is unverified, *THEN* do not treat PSNR/SSIM as primary evidence; rely on `P` via FID/KID and add `S`/`D` as required.
2. *IF* claiming semantic or structural preservation, *THEN* report `S` with at least one test that detects object removal/hallucination on critical classes.
3. *IF* claiming benefit for detection/segmentation, *THEN* report `D` on *real* nighttime benchmarks and include the no-translation baseline.

Table 7 Minimum evidence checklist for IDT claims. Evidence codes follow Table 6: P (perceptual/distributional), S (semantic/structural), D (downstream utility), T (temporal stability)

Claim/use-case	Mandatory evidence	Minimum tests (acceptance-grade)	Suitable benchmark types
"Looks realistic" translation demo	P	FID/KID; LPIPS only when correspondence is justified; include failure cases (small objects, glare, under-exposure).	Large unpaired day/night sets; paired or pseudo-paired subsets for diagnostics.
"Preserves scene content / structure"	S (+P)	Label/task-network consistency where labels exist; object retention & hallucination counts for critical classes; geometry/matchability tests for localisation/retrieval.	Benchmarks with labels/correspondence (e.g. driving segmentation, localisation datasets).
"Improves downstream perception" (det/seg)	D (+S)	mAP/mIoU on *real night* test sets; report deltas vs no-translation; prefer ≥ 2 downstream models or architectures.	Night detection/segmentation benchmarks; multi-condition driving suites with night splits.
"Video/streaming ready"	T (+D if claimed)	Warp-consistency, drift/stability over long horizons; explicit night stressors (headlights/exposure); downstream tracking/localisation over sequences when claimed.	Night video sequences (driving/surveillance), long-horizon repeats.
"Improves localisation / retrieval"	D (+S)	Recall@k, pose error, keypoint match statistics; evaluate matchability under illumination change rather than photorealism.	Day-night localisation/retrieval benchmarks (repeat traversals).
"Robust under glare/noise/weather"	S+D (stratified)	Report factor-stratified results (glare/noise/weather bins) and worst-case tails; include diagnostic subsets with controlled stressors.	Multi-condition benchmarks; adverse-condition suites with night subsets.

Table 8 Minimum acceptable experiment packages for operationalising the P/S/D/T protocol. Authors may exceed these packages, but papers that fall substantially below them should narrow their claims accordingly

Paper type	Recommended minimum datasets	Recommended metric combination	Mandatory stress tests/reporting
Image-based IDT	One correspondence or structure supporting benchmark plus one real night downstream benchmark. A defensible minimum is Dark Zurich for S plus ACDC, NightCity, Nighttime Driving, or ExDark for D, depending on task; add CROWD for broad held-out geographic and temporal robustness audits when pseudo labels are acceptable.	Report P with FID+KID and LPIPS only when correspondence assumptions are explicit; report at least one S check, for example mIoU, boundary retention, object retention, or hallucination sensitive analysis, and one D result on real night data with the no translation baseline.	Include worst case slices for small or distant objects, deep shadows, glare or saturation, and mixed artificial lighting. If robustness is claimed, add held out domain or cross dataset transfer. State synthetic to real use explicitly when applicable.
Video-based IDT	At least one real night video benchmark with task or tracking labels, such as BDD100K (night), NightOwls, or CROWD when geographic breadth and routine continuous driving are central; use additional sequence data when localisation or retrieval is the target.	Report frame level P only as supporting evidence, together with at least one D task metric on real night video and one T measure such as warping consistency, drift or flicker statistics, or sequence level downstream stability. Add S where structure labels or correspondences are available.	Test moving headlights, specular sweeps, exposure shifts, motion blur, and fast camera or object motion. Report whether inference is causal or uses future context, and state latency or throughput, hardware, and input resolution for any online claim.

4. *IF* the paper motivates safety-relevant deployment, *THEN* include a worst-case analysis (small objects, glare, deep shadows) and do not summarise solely with mean image-level scores.
5. *IF* the method targets video or streaming use, *THEN* report T and include night-specific temporal stressors (headlights, exposure changes, specular sweeps).
6. *IF* reporting LPIPS, *THEN* state what correspondence assumption is used (true pairs, pseudo-pairs, same-location repeats) and how dynamic objects are handled.
7. *IF* using synthetic night for training or evaluation, *THEN* validate claims with D on genuine night data and report the synthetic-to-real gap explicitly.
8. *IF* claiming robustness across locations, sensors, or collection conditions, *THEN* report cross-dataset or held-out-domain performance rather than relying on a single in-domain benchmark.
9. *IF* the method claims to address glare/noise/weather, *THEN* stratify results by severity proxies (e.g.

Table 9 Evidence-aligned comparison digest across representative IDT settings. The table is not a unified leaderboard: it indicates which families and quantitative outcomes are commonly reported and what comparisons are methodologically defensible under heterogeneous protocols

Setting/ benchmark type	Representative families	Common quantitative evidence	What comparison is usually valid	Main limitation
Unpaired day-night image translation (e.g. Dark Zurich / ACDC style settings)	Cycle or identity GANs; semantic anchors; correspondence variants	FID / LPIPS; sometimes mIoU or detection transfer on translated or real night data	Within-setting comparison under matched splits, supervision, and task model	Absolute cross-paper ranking is weak when training data and task protocols differ
Task-conditioned localisation / retrieval	Task-coupled translation and condition-robust localisation families	Recall@k, localisation success, pose error, retrieval accuracy	Comparisons using the same retrieval dataset and task definition	Perceptual metrics are secondary and do not establish task robustness
Video IDT	Temporal GANs; video translators; sequence-constrained methods	Warping error, flicker measures, tracking or segmentation stability	Comparisons with matched sequence protocol, frame rate, and causal assumptions	Long-horizon drift and memory or latency costs are often underreported
Diffusion / foundation-model pipelines	Pretrained diffusion priors; one-step diffusion translators	Usually FID / LPIPS; occasionally downstream transfer or structure checks	Family-level comparison of evidence mix and within-paper ablations	Realism gains are not sufficient to support deployment claims without S/D/T evidence

Table 11 Aggregate artefact availability across the audited 34 methods (snapshot: December 19, 2025). "Recipe" corresponds to a runnable compendium (configs/hparams + environment + commands); "Licence" requires explicit terms for code (and, when stated, weights)

Signal	Count	Share
Code	29 / 34	85.3%
Pretrained weights	22 / 34	64.7%
Licence terms	21 / 34	61.8%
Reproducibility recipe	19 / 34	55.9%

saturation/under-exposure masks, noise estimates, weather tags) and report tail performance.

10. *IF* semantic/task networks are used as constraints, *THEN* report ablations that isolate the constraint effect (e.g. frozen vs fine-tuned, with/without the task loss).
11. *IF* downstream utility is reported, *THEN* prefer at least two downstream models/architectures to reduce model-specific bias.
12. *IF* a method is positioned for adoption, *THEN* provide sufficient artefacts (recipe + licence) so results can be verified under the protocol used.
13. *IF* robustness across locations/sensors is claimed, *THEN* cross-dataset performance must be reported (for example, Dark Zurich → ACDC or BDD100K Night → NightOwls); otherwise, the claim should be treated as an over-claim.
14. *IF* a method is proposed for online, streaming, or safety-relevant deployment, *THEN* report inference latency or throughput, input resolution, hardware, and whether processing is causal, batched, or uses future context; otherwise, real-time applicability should be treated as unsubstantiated.*Minimum acceptable experiment packages*. To make the checklist more directly operational for authors, Table 8 summarises compact minimum experiment packages for image-based and video-based IDT papers. These packages are not intended as universal templates for every task, but as acceptance-grade

Table 10 Artefact availability matrix for selected methods referenced in this survey

Paper/Method	Focus	Code	Wts.	Data	Rec.	Lic.
AU-GAN [88]	Adverse condition translation (incl. N2D)	✓	✓	✓	✓	–
CycleGAN [18]	Unpaired translation baseline	✓	✓	✓	✓	✓
Daydriex [19]	N2D driving pipeline	–	–	–	–	–
DiCo [29]	N2D for surveillance	–	–	–	–	–
DUNIT [53]	Instance aware translation	✓	–	✓	✓	–
$N2D^3$ [58]	Degradation disentanglement	–	–	✓	–	–
Paired-N2D [30]	Supervised N2D (synthetic pairs)	✓	–	✓	–	–
Pix2Pix [17]	Paired translation baseline	✓	✓	✓	✓	✓
SPN2D-GAN [89]	Semantic prior N2D	✓	–	✓	–	–
ToDayGAN [40]	N2D for localisation	✓	✓	✓	✓	✓
UNIT [61]	Shared latent baseline	✓	–	✓	✓	–

"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

baselines that make it harder to over-claim from perceptual metrics alone.

Taken together, these rules define the comparison framework used in this survey: methods are not ranked by isolated headline scores, but compared according to whether their claims are supported by evidence that is valid for the supervision regime, benchmark setting, and deployment context under consideration.

7.3 Evidence-aligned qualitative synthesis

Table 4 and Table 9 should be read together: the former summarises family-level strengths and weaknesses, while the latter shows where quantitative comparison is actually defensible across representative IDT settings.

This survey does not introduce new training runs or a formal meta-analysis over heterogeneous protocols. Instead, it compares papers using supervision regime, night-domain gap factors, constraint family, and evidence type under P/S/D/T. That choice is deliberate: training data, evaluation splits, hardware assumptions, and metric definitions vary enough across IDT papers that pooled score aggregation would create a misleading sense of commensurability. We therefore treat within-paper ablations, matched-task comparisons, and evidence-aligned reporting as stronger bases for interpretation than raw cross-paper score ranking, and we use the synthesis to identify where reported strengths remain limited by weak evidence, fragile comparability, unresolved failure modes, or incomplete reproducibility support.

This framing also sharpens scope and validation. Enhancement-only and adaptation-only results are not treated as directly commensurate with IDT validation unless translated outputs themselves are evaluated under the survey protocol. Table 9 therefore summarises representative IDT settings, the quantitative evidence commonly reported, and the comparisons that remain methodologically defensible across method families without imposing a misleading leaderboard.

The practical value of the protocol is clearest in worked comparisons across claim types. A paper that reports strong FID or LPIPS improvements but no object-retention test remains weakly supported for semantic preservation, because P alone does not exclude hallucinated structure or small-object deletion. By contrast, a method that combines acceptable perceptual quality with downstream gains on real nighttime detection or segmentation benchmarks provides stronger support for deployment-relevant utility under D. Similarly, for video translation, perceptually plausible frame-wise outputs remain insufficient unless temporal consistency is supported by T, for example through warping-based stability checks or sequence-level downstream evaluation. These comparisons do not create a unified leaderboard, but they do show how the protocol changes the interpretation of otherwise similar headline claims.

Several recurring cross-paper contrasts make this point more tangible. First, classic unpaired GAN papers often appear strong under perceptual reporting because they improve day-night realism or distributional similarity, yet many provide little direct evidence that small objects, lane markings, or traffic signs are preserved under low-evidence conditions. Under our protocol, such papers remain primarily P-supported unless they add explicit structure-aware or task-aware validation. Second, task-coupled methods can sometimes look less impressive in image-only comparisons while still being more persuasive for deployment relevance if they report improvements on matched downstream localisation, detection, or segmentation tasks under real night conditions. Third, in video settings, frame-wise visual quality can make a method look competitive even when temporal evidence is weak; once T is required, methods without warping-based consistency checks, drift analysis, or sequence-level downstream evaluation become much less convincing for streaming use. Finally, recent diffusion-backed translators often benefit from strong generative priors and can therefore look highly competitive under realism metrics, but in the absence of explicit S, D, or T evidence they should still be interpreted as evidence-incomplete for safety-relevant deployment.

Among unpaired methods, the main analytical contrast is not simply GAN versus non-GAN, but weakly anchored distribution alignment versus explicitly content-anchored translation. Cycle/identity-based baselines can improve global appearance and reduce coarse day-night mismatch, but in low-evidence nighttime regions they remain vulnerable to semantic drift, small-object removal, and hallucinated structure because alignment is enforced at the domain level rather than at the object or task level. By contrast, semantic- or instance-anchored methods more directly protect object identity and scene layout, especially for roads, vehicles, and traffic signs, but they inherit the limitations of the task networks or labels used as anchors and may overfit to a narrow downstream notion of correctness.

A second cross-family contrast appears between correspondence-based and physics-informed families. Correspondence or contrastive constraints are often effective when local structure remains partially visible, because they suppress texture drift and preserve patch-level consistency without requiring strong image-formation assumptions. Physics-informed and decomposition-based approaches are better aligned with night-specific degradations such as under-exposure, glare, and noise, but their gains depend on how well the assumed reflectance, illumination, or

degradation model matches real outdoor conditions. In practice, correspondence-based methods tend to fail when evidence is too weak to support stable matching, whereas physics-based methods tend to fail when mixed lighting, specularities, or non-Lambertian effects violate the underlying assumptions.

Recent diffusion and foundation-model pipelines sharpen a third analytical tension: realism has improved faster than deployment-grade faithfulness. Strong generative priors can produce highly plausible outputs and better coverage of rare appearances, but the literature still provides limited quantitative evidence that these gains translate into reliable preservation of geometry, object integrity, or downstream task utility under night-specific stressors. As a result, diffusion-backed methods should currently be interpreted as promising but evidence-incomplete for safety-critical use, unless they are supported by explicit `S`, `D`, and, where relevant, `T` evaluation.

Overall, the literature suggests that the central trade-off is not realism versus fidelity in the abstract, but weakly constrained realism versus evidence-supported preservation under night-specific ambiguity.

7.4 Artefact matrix and implications

Table 10 reports the method-level availability matrix under the rubric in Table 3. The matrix is intended as a transparency aid: it states which resources a reader can reasonably expect to find *at the snapshot date*. It is not a judgement of scientific merit. A method can be technically strong even if artefacts are not publicly released, but then verification and comparison costs shift to the reader and practical adoption is often blocked.

For visibility in the main paper, Table 10 is presented here as a soft colour-coded availability matrix. To remain readable in print and not rely on colour alone, the matrix retains explicit symbols in each cell.

7.4.1 Aggregate availability and "reproducibility readiness"

To make the audit actionable, we summarise aggregate availability signals in Table 11. Two patterns are particularly relevant for practitioners. First, the main bottleneck is rarely *source code* alone: code availability is comparatively high, but complete *recipes* (configs + environments + runnable commands) and explicit *licences* are less consistently available. Second, pretrained weights reduce reproduction cost substantially, but weights without an explicit recipe and dataset split specification can still prevent fair comparison under a shared protocol.

Readiness tiers. For practical interpretation, it is useful to group methods by "reproducibility readiness" rather than by a single signal: We use these readiness tiers as a practical reproducibility scoring scheme for the survey: R0 indicates minimal reproducibility support, R1 indicates partial but non-trivial reproducibility support, and R2 indicates research-compendium-level support sufficient for realistic rerun and comparison under a shared protocol. Operationally, R2 requires public code, pretrained weights or checkpoints where applicable, runnable training or evaluation recipes, explicit dataset split information, and clear licensing terms.

- **Tier R0 (link-only).** Code is missing or incomplete, and/or no weights/recipe are provided; reproduction requires substantial re-implementation effort.
- **Tier R1 (runnable with effort).** Code exists and either weights or partial recipe is provided; reproduction is possible but requires non-trivial reverse engineering (e.g. missing splits, missing environment pins).
- **Tier R2 (research-compendium).** Code, weights, and a runnable recipe (plus explicit licences) are provided; reproduction and fair comparison under a shared protocol are realistically feasible.

7.4.2 Implications for evidence claims and fair comparison

The audit has three direct implications for how readers should interpret the empirical literature. Because this survey does not run repository-level verification experiments, the audit should be interpreted as a reproducibility-readiness assessment rather than an executability certificate.

Condition comparisons on artefact completeness. When a paper provides only image examples or only a partial release (e.g. code without configs), the reported gains are difficult to validate under the P/S/D/T protocol. For evidence synthesis and "best practice" recommendations, we therefore treat results from Tier R2 releases as higher-weight evidence for *replicable* performance, especially when the paper makes deployment or safety-relevant claims.

Treat licences as a first-class deployment constraint. In safety-relevant outdoor systems, unclear or missing licences for code or pretrained weights can block adoption even when a method is technically promising. We therefore record licence availability explicitly rather than assuming common defaults.

Separate availability from executability. A repository can be public yet non-executable due to dependency drift or undocumented preprocessing. We therefore recommend pairing public releases with pinned dependencies (e.g. environment files or containers) and a minimal end-to-end script

that reproduces at least one headline result under the stated protocol.

We recommend that future IDT work releases, at minimum, a versioned repository with training and evaluation scripts, exact dataset splits, pinned dependencies, and explicit licences for both code and pretrained weights [90–92].

- *Supported*. Public code and pretrained weights are increasingly common, but missing licences and incomplete run recipes still frequently block practical adoption and fair comparison.
- *Promising*. Release of full research compendia (versioned code, splits, configs, evaluation scripts) can materially improve verification and reduce hidden implementation variance in generative pipelines.
- *Unclear*. Snapshot availability alone is insufficient to claim reproducibility; stronger meta-scientific evidence requires at least limited execution/validation subsets and a structured rubric.

8 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.

8.1 Computational complexity and deployment feasibility

Computational cost is an important but inconsistently reported dimension of outdoor IDT. Direct latency or throughput comparisons across papers are often not methodologically valid because hardware, image resolution, sequence length, batching strategy, and use of offline or future-frame context differ substantially. Nevertheless, family-level tendencies are clear. Cycle- and identity-based GAN translators are often lighter at inference than diffusion-backed pipelines, although semantic- or task-anchored variants can incur additional cost through auxiliary networks. Physics-informed and decomposition-based approaches vary widely, ranging from relatively lightweight feed-forward models to more expensive iterative or multi-branch pipelines. Video methods typically impose the highest deployment burden because temporal buffering, optical flow, spatio-temporal discriminators, or sequence-level feature fusion increase both memory use and latency. For strict real-time or safety-relevant deployment, claims of practical feasibility should therefore be supported by explicit reporting of hardware, resolution, throughput or latency, and whether the method operates causally or depends on future context.

With these comparative and evaluation issues in place, we now turn to the broader implications for responsible use, reproducibility, and future work.

8.2 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 [8, 20]. 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 [23, 25]. 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 [58, 64], while semantic and instance constraints aim to anchor translation to task-relevant structure [53, 63]. 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 [39, 79].

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.

8.3 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 [5, 79]. Adverse weather further compounds the appearance gap and interacts with night illumination in ways that are not captured by clear-night benchmarks [79].

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. Large geographically distributed dashcam resources such as CROWD can support this goal by enabling held-out locality, country, and continent splits for routine urban driving, although their machine-generated annotations should be treated as pseudo labels rather than manual ground truth [80]. 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.

8.4 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 [26, 68]. 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 [72, 73].

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 [26, 68]. For flow-based methods, reports should include sensitivity analyses in flow failure modes typical of nighttime video [72, 73].

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 5 provides a compact dataset shortlist for acceptance-grade evidence.

Appendix

Full reference tables

To improve readability in the main narrative, the full compendium tables are placed here. The main paper contains condensed, decision-oriented tables and figures; this appendix provides the complete crosswalks for reproducibility and lookup.

Complete method-to-constraint crosswalk

The full method-level crosswalk referenced throughout the paper is given in Table 12.

Complete dataset catalog

Table 13 lists the full dataset catalog referenced in Section 7.

Table 12 Representative crosswalk from methods to constraints for IDT

Method	Yr	Venue	Sup.	Gap	Constraints					Primary eval.			
					Cyc/Id	Sem/Inst	Corr.	Phys.	Temp.	P	S	D	T
CycleGAN [18]	2017	ICCV	Unpaired	I	✓	–	–	–	–	✓	–	–	–
DualGAN [60]	2017	ICCV	Unpaired	I	✓	–	–	–	–	✓	–	–	–
Pix2Pix [17]	2017	CVPR	Paired	I	–	–	–	–	–	✓	✓	–	–
UNIT [61]	2017	NIPS	Unpaired	I	✓	–	–	–	–	✓	–	–	–
vid2vid [26]	2018	NeurIPS	P. (vid)	I,M	–	–	–	–	✓	✓	–	–	✓
MoCycleGAN [62]	2019	arXiv	Unp. (vid)	I,M	✓	–	–	–	✓	✓	–	–	✓
ToDayGAN [40]	2019	ICRA	Unp./task	I	✓	✓	–	–	–	–	–	✓	–
CUT [65]	2020	ECCV	Unpaired	I	–	–	✓	–	–	✓	–	–	–
DUNIT [53]	2020	CVPR	Unp./inst.	I	✓	✓	–	–	–	–	✓	✓	–
ForkGAN [93]	2020	ECCV	Unpaired	I,W	–	–	–	✓	–	✓	–	✓	–
Night-to-Day (Online) [31]	2020	T-IV	Unp. (vid)/task	I,M	–	✓	–	–	✓	–	–	✓	–
AU-GAN [88]	2021	arXiv	Unpaired	I,W	✓	–	–	–	–	✓	–	–	–
CoMoGAN [94]	2021	CVPR	Unpaired	I	–	–	–	✓	–	✓	–	–	–
Daydriex [19]	2021	arXiv	Unp./aux	I	✓	–	✓	–	–	✓	–	✓	–
D2N ISP Synthesis [57]	2022	CVPR	P. (synth)	I,N	–	–	–	✓	–	–	–	✓	–
ManiFest [95]	2022	ECCV	Few-shot/unp.	I	–	–	–	–	–	✓	–	–	–
N2D-GAN [96]	2022	ICME	Unp./weak	I	✓	✓	–	–	–	✓	–	–	–
PITI [24]	2022	arXiv	P./Unp.	I	–	–	–	–	–	✓	–	–	–
SPN2D-GAN [89]	2022	TMM	Unp./sem	I	✓	✓	–	–	–	–	✓	–	–
2PCNet [97]	2023	CVPR	Unp./task	I,G,N	–	✓	–	–	–	–	–	✓	–
Dark Side Augmentation [67]	2023	ICCV	Unp./task	I	–	✓	–	–	–	–	–	✓	–
DiCo [29]	2023	arXiv	Unpaired	I	–	–	✓	–	–	✓	✓	–	–
RefN2D-Guide GAN [98]	2023	LNCS	Unp./ref+sem	I	–	✓	✓	–	–	✓	✓	–	–
UVCGAN [54]	2023	WACV	Unpaired	I	✓	–	–	–	–	✓	–	–	–
CycleGAN-Turbo [99]	2024	arXiv	Unpaired	I,W	–	–	–	–	–	✓	–	–	–
$N2D^3$ [58]	2024	arXiv	Unpaired	I,G,N	–	–	✓	✓	–	✓	✓	–	–
Paired-N2D [30]	2024	IEEE Access	P. (synth)	I	–	–	–	–	–	✓	✓	–	–
pix2pix-Turbo [99]	2024	arXiv	Paired	I,W	–	–	–	–	–	✓	–	–	–
SGA-D2N [56]	2024	Sensors	Unp./sem+geom	I	✓	✓	–	–	–	✓	–	–	–
RLA-Training [64]	2025	Math.	Unp.+p./task	I	✓	–	–	✓	–	–	–	✓	–
Bridging Day and Night [59]	2026	AAAI	Unp./sem	I	–	✓	–	–	–	✓	✓	✓	–

Sup. abbreviations: P. = paired, Unp. = unpaired. Gap codes: I = illumination, G = glare, N = noise, W = weather, M = motion or video. Constraint codes: Cyc/Id = cycle and identity, Sem/Inst = semantic, instance, or task-network constraints, Corr. = contrastive or correspondence, Phys. = physics or decomposition, Temp. = temporal. Evaluation codes: P = perceptual/distributional, S = semantic/structural, D = downstream utility, T = temporal stability

Table 13 Representative datasets used in day and night outdoor vision

Dataset (ref.)	Platform	Pairing	Labels	Sensors	Protocol codes	Most defensible evaluation
Aachen Day-Night [77]	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.
ACDC [79]	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.
BDD100K [7]	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.
CROWD [80]	Driving (video)	Unpaired collection with segment-level day/night labels	Machine-generated COCO boxes and segment-local tracks; vehicle type and locality metadata	RGB	D T	External robustness and temporal-continuity audits: detector/tracker consistency, stratified day/night evaluation, and held-out locality/country/continent tests; avoid pixel-aligned metrics because day-night pairs are not provided.
D^2-City [100]	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).
Dark Zurich [39]	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.
ExDark [81]	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).
Gardens Point Walking [101]	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.
MSLS (Mapillary Street-Level Sequences) [102]	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.
NightCity [83]	Driving (images)	Unpaired collection	Fine semantic masks	RGB	S D	Segmentation: mIoU; report sensitivity to exposure and illumination modelling.
NightOwls [8]	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.
Nighttime Driving [38]	Driving (images)	Unpaired collection	Coarse semantic masks	RGB	S D	Segmentation: mIoU; adaptation gains measured on real night, not only translated imagery.
RobotCar Seasons [77]	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.
Tokyo 24/7 [84]	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.
UAVDark135 [103]	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.

"Pairing" denotes correspondences across conditions or modalities. Protocol codes follow Table 6. The evaluation column prioritises task aligned evidence rather than image realism metrics alone

Author contributions M.S.A. conceived the study, designed the survey scope and taxonomy, conducted the literature review and artefact audit, and wrote the main manuscript. P.S. contributed to the literature collection and screening, assisted in curating the method/dataset comparisons and tables, and reviewed and edited the manuscript. P.B. provided technical guidance on evaluation criteria and application framing, critically revised the manuscript for clarity and completeness, and supervised the overall work. All authors reviewed and approved the final manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ma, X., Ouyang, W., Simonelli, A., Ricci, E.: 3d object detection from images for autonomous driving: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **46**(5), 3537–3556 (2024). https://doi.org/10.1109/TPAMI.2023.3346386
2. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. Int. J. Comput. Vision **128**(2), 261–318 (2020). https://doi.org/10.1007/s11263-019-01247-4
3. Du, H., Shi, H., Zeng, D., Zhang, X.-P., Mei, T.: The elements of end-to-end deep face recognition: A survey of recent advances. ACM computing surveys (CSUR) **54**(10s), 1–42 (2022). https://doi.org/10.1145/3507902
4. Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.-M., Gu, J., Loy, C.C.: Low-light image and video enhancement using deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(12), 9396–9416 (2021). https://doi.org/10.1109/TPAMI.2021.3126387
5. Liu, J., Xu, D., Yang, W., Fan, M., Huang, H.: Benchmarking low-light image enhancement and beyond. Int. J. Comput. Vision **129**(4), 1153–1184 (2021). https://doi.org/10.1007/s11263-020-01418-8
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016). https://doi.org/10.48550/arXiv.1604.01685
7. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2636–2645 (2020). https://doi.org/10.1109/cvpr42600.2020.00271
8. Neumann, L., Karg, M., Zhang, S., Scharfenberger, C., Piegert, E., Mistr, S., Prokofyeva, O., Thiel, R., Vedaldi, A., Zisserman, A., *et al.*: Nightowls: A pedestrians at night dataset. In: Asian Conference on Computer Vision, pp. 691–705 (2018). https://doi.org/10.1007/978-3-030-20887-5_43 . Springer
9. Wang, G., Ma, Y., Huang, J., Fan, F., Li, H., Li, Z.: Instance segmentation of pigs in infrared images based on inpc model. Infrared Physics & Technology **141**, 105491 (2024). https://doi.org/10.1016/j.infrared.2024.105491
10. Wang, G., Ma, Y., Huang, J., Fan, F., Wang, Z.: Measurement of pig body temperature based on ear segmentation and multi-factor infrared temperature compensation. IEEE Trans. Instrum. Meas. (2025). https://doi.org/10.1109/TIM.2025.3538062
11. Wang, G., Ma, Y., Huang, J., Wu, K., Tang, L., Cai, Z., Fan, F.: Clte: Infrared and visible image fusion through cross-domain learning and texture-contrast enhancement. Infrared Physics & Technology, 106230 (2025) https://doi.org/10.1016/j.infrared.2025.106230
12. Lu, Y., Huang, J., Ma, Y., Fan, F., Wu, K., Wang, G.: Frequency-aware retinex-wavelet decomposition hybrid network and luminance-guided transformers for low-light image enhancement. Optics & Laser Technology **194**, 114432 (2026). https://doi.org/10.1016/j.optlastec.2025.114432
13. Jin, Z., Qiu, Y., Zhang, K., Li, H., Luo, W.: Mb-taylorformer v2: Improved multi-branch linear transformer expanded by taylor formula for image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence 47(7), 5990–6005 (2025) https://doi.org/10.1109/TPAMI.2025.3559891
14. Zhang, K., Luo, W., Zhong, Y., Ma, L., Liu, W., Li, H.: Adversarial spatio-temporal learning for video deblurring. IEEE Trans. Image Process. **28**(1), 291–301 (2019). https://doi.org/10.1109/TIP.2018.2867733
15. Zhang, K., Wang, T., Luo, W., Ren, W., Stenger, B., Liu, W., Li, H., Yang, M.-H.: Mc-blur: A comprehensive benchmark for image deblurring. IEEE Trans. Circuits Syst. Video Technol. **34**(5), 3755–3767 (2024). https://doi.org/10.1109/TCSVT.2023.3319330
16. Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: Methods and applications. IEEE Trans. Multimedia **24**, 3859–3881 (2021). https://doi.org/10.1109/TMM.2021.3109419
17. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017). https://doi.org/10.1109/CVPR.2017.632
18. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017). https://doi.org/10.1109/ICCV.2017.244
19. Lee, E., Kang, S.: Daydriex: Translating nighttime scenes towards daytime driving experience at night. Appl. Sci. **11**(5), 2013 (2021). https://doi.org/10.3390/app11052013
20. Jobson, D.J., Rahman, Z., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Trans. Image Process. **6**(7), 965–976 (1997). https://doi.org/10.1109/83.597272
21. Gehler, P.V., Rother, C., Blake, A., Minka, T., Sharp, T.: Bayesian color constancy revisited. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008). https://doi.org/10.1109/CVPR.2008.4587765 . IEEE

22. Land, E.H., McCann, J.J.: Lightness and retinex theory. J. Opt. Soc. Am. **61**(1), 1–11 (1971). https://doi.org/10.1364/JOSA.61.000001
23. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6228–6237 (2018). https://doi.org/10.1109/CVPR.2018.00652
24. Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F.: Pretraining is all you need for image-to-image translation. arXiv preprint arXiv:2205.12952 (2022) https://doi.org/10.48550/arXiv.2205.12952
25. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
26. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: Conference on Neural Information Processing Systems (NeurIPS) (2018). https://doi.org/10.48550/arXiv.1808.06601
27. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Comput. Graphics Appl. **21**(5), 34–41 (2002). https://doi.org/10.1109/38.946629
28. Lengyel, A., Garg, S., Milford, M., Gemert, J.C.: Zero-shot day-night domain adaptation with a physics prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4399–4409 (2021). https://doi.org/10.1109/ICCV48922.2021.00436
29. Lan, G., Zhao, B., Li, X.: Disentangled contrastive image translation for nighttime surveillance. arXiv preprint arXiv:2307.05038 (2023) https://doi.org/10.48550/arXiv.2307.05038
30. Lakmal, H., Dissanayake, M.B., Aramvith, S.: Light the way: an enhanced generative adversarial network framework for night-to-day image translation with improved quality. IEEE Access **12**, 165963–165978 (2024). https://doi.org/10.1109/ACCESS.2024.3491792
31. Schutera, M., Hussein, M., Abhau, J., Mikut, R., Reischl, M.: 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 (2020). https://doi.org/10.1109/TIV.2020.3039456
32. Yang, S., Sun, M., Lou, X., Yang, H., Zhou, H.: An unpaired thermal infrared image translation method using gma-cyclegan. Remote Sensing **15**(3), 663 (2023). https://doi.org/10.3390/rs15030663
33. Guo, J., Ma, J., García-Fernández, Á.F., Zhang, Y., Liang, H.: A survey on image enhancement for low-light images. Heliyon **9**(4) (2023) https://doi.org/10.1016/j.heliyon.2023.e14558
34. Huang, S., Jin, X., Jiang, Q., Liu, L.: Deep learning for image colorization: Current and future prospects. Eng. Appl. Artif. Intell. **114**, 105006 (2022). https://doi.org/10.1016/j.engappai.2022.105006
35. Anwar, S., Tahir, M., Li, C., Mian, A., Khan, F.S., Muzaffar, A.W.: Image colorization: A survey and dataset. Information Fusion **114**, 102720 (2025). https://doi.org/10.1016/j.inffus.2024.102720
36. Hoyez, H., Schockaert, C., Rambach, J., Mirbach, B., Stricker, D.: Unsupervised image-to-image translation: A review. Sensors **22**(21), 8540 (2022). https://doi.org/10.3390/s22218540
37. Saxena, S., Teli, M.N.: Comparison and analysis of image-to-image generative adversarial networks: a survey. arXiv preprint arXiv:2112.12625 (2021) https://doi.org/10.48550/arXiv.2112.12625
38. Dai, D., Van Gool, L.: Dark model adaptation: Semantic image segmentation from daytime to nighttime. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3819–3824. IEEE Press, Piscataway, NJ, USA (2018). https://doi.org/10.1109/ITSC.2018.8569387
39. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7374–7383 (2019). https://doi.org/10.1109/ICCV.2019.00747
40. Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., Van Gool, L.: Night-to-day image translation for retrieval-based localization. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 5958–5964 (2019). https://doi.org/10.1109/ICRA.2019.8794387 . IEEE
41. Cherian, A., Sullivan, A.: Sem-gan: Semantically-consistent image-to-image translation. In: 2019 Ieee Winter Conference on Applications of Computer Vision (wacv), pp. 1797–1806 (2019). https://doi.org/10.1109/WACV.2019.00196 . IEEE
42. Huang, X., Liu, M.-Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–189 (2018). https://doi.org/10.1007/978-3-030-01219-9_11
43. Pitié, F.: Advances in colour transfer. IET Comput. Vision **14**(6), 304–322 (2020). https://doi.org/10.1049/iet-cvi.2019.0920
44. Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. *Proceedings of Machine Learning Research***80**, 1989–1998 (2018) https://doi.org/10.48550/arXiv.1711.03213
45. Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 465–476. Curran Associates Inc., Red Hook, NY, USA (2017). https://doi.org/10.5555/3294771.3294816
46. Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–51 (2018). https://doi.org/10.1007/978-3-030-01246-5_3
47. Zuiderveld, K.J.: Contrast limited adaptive histogram equalization. In: Graphics Gems (1994). https://doi.org/10.1016/B978-0-12-336156-1.50061-6
48. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 257–266 (2002). https://doi.org/10.1145/566654.566574
49. Barrow, H., Tenenbaum, J., Hanson, A., Riseman, E.: Recovering intrinsic scene characteristics. Comput. vis. syst **2**(3–26), 2 (1978). https://doi.org/10.5555/2968618.2968788
50. Barron, J.T., Malik, J.: Color constancy, intrinsic images, and shape estimation. In: European Conference on Computer Vision, pp. 57–70 (2012). https://doi.org/10.1007/978-3-642-33765-9_5 . Springer
51. Matsushita, Y., Nishino, K., Ikeuchi, K., Sakauchi, M.: Illumination normalization with time-dependent intrinsic images for video surveillance. IEEE Trans. Pattern Anal. Mach. Intell. **26**(10), 1336–1347 (2004). https://doi.org/10.1109/TPAMI.2004.86
52. Toet, A., Hogervorst, M.A.: Progress in color night vision. Opt. Eng. **51**(1), 010901–010901 (2012). https://doi.org/10.1117/1.OE.51.1.010901
53. Bhattacharjee, D., Kim, S., Vizier, G., Salzmann, M.: Dunit: Detection-based unsupervised image-to-image translation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4786–4795 (2020). https://doi.org/10.1109/CVPR42600.2020.00484
54. Torbunov, D., Huang, Y., Yu, H., Huang, J., Yoo, S., Lin, M., Viren, B., Ren, Y.: Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of

Computer Vision, pp. 702–712 (2023). https://doi.org/10.1109/WACV56688.2023.00077
55. Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2337–2346 (2019). https://doi.org/10.1109/CVPR.2019.00244
56. Bang, G., Lee, J., Endo, Y., Nishimori, T., Nakao, K., Kamijo, S.: Semantic and geometric-aware day-to-night image translation network. Sensors **24**(4), 1339 (2024). https://doi.org/10.3390/s24041339
57. Punnappurath, A., Abuolaim, A., Abdelhamed, A., Levinshtein, A., Brown, M.S.: Day-to-night image synthesis for training nighttime neural isps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10769–10778 (2022). https://doi.org/10.1109/CVPR52688.2022.01050
58. Lan, G., Yang, Y., Wang, Z., Wang, D., Zhao, B., Li, X.: Night-to-day translation via illumination degradation disentanglement. arXiv preprint arXiv:2411.14504 (2024) https://doi.org/10.48550/arXiv.2411.14504
59. Li, S., Tan, L., Tan, R.T.: Bridging day and night: Target-class hallucination suppression in unpaired image translation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, pp. 6424–6432 (2026). https://doi.org/10.1609/aaai.v40i8.37570
60. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2849–2857 (2017). https://doi.org/10.1109/ICCV.2017.310
61. Liu, M.-Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 700–708. Curran Associates Inc., Red Hook, NY, USA (2017). https://doi.org/10.5555/3294771.3294838
62. Chen, Y., Pan, Y., Yao, T., Tian, X., Mei, T.: Mocycle-gan: Unpaired video-to-video translation. In: Proceedings of the 27th ACM International Conference on Multimedia. MM '19, pp. 647–655. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3343031.3350937
63. Shiotsuka, D., Lee, J., Endo, Y., Javanmardi, E., Takahashi, K., Nakao, K., Kamijo, S.: Gan-based semantic-aware translation for day-to-night images. In: 2022 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–6 (2022). https://doi.org/10.1109/ICCE53296.2022.9730532 . IEEE
64. Lee, Y.-J., Go, Y.-H., Lee, S.-H., Son, D.-M., Lee, S.-H.: Night-to-day image translation with road light attention training for traffic information detection. Mathematics **13**(18), 2998 (2025). https://doi.org/10.3390/math13182998
65. Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX, pp. 319–345. Springer, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58545-7_19
66. Jia, Z., Yuan, B., Wang, K., Wu, H., Clifford, D., Yuan, Z., Su, H.: Semantically robust unpaired image translation for data with unmatched semantics statistics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14273–14283 (2021). https://doi.org/10.1109/ICCV48922.2021.01401
67. Mohwald, A., Jenicek, T., Chum, O.: Dark side augmentation: Generating diverse night examples for metric learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11153–11163 (2023). https://doi.org/10.1109/ICCV51070.2023.01024
68. Wei, X., Zhu, J., Feng, S., Su, H.: Video-to-video translation with global temporal consistency. In: Proceedings of the 26th ACM International Conference on Multimedia. MM '18, pp. 18–25. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3240508.3240708
69. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Fresco: Spatial-temporal correspondence for zero-shot video translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8703–8712 (2024). https://doi.org/10.1109/CVPR52733.2024.00831
70. Taufiq, M.F.R., Rahadianti, L.: Cyclegan for day-to-night image translation: a comparative study. IAES International Journal of Artificial Intelligence 14(3), 2347–2357 (2025) https://doi.org/10.11591/ijai.v14.i3.pp2347-2357
71. Liu, H., Cheng, H., Ye, L.: Dnit: enhancing day-night image-to-image translation through fine-grained feature handling (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 23563–23564 (2024). https://doi.org/10.1609/aaai.v38i21.30474
72. Ruder, M., Dosovitskiy, A., Brox, T.: Artistic style transfer for videos. In: German Conference on Pattern Recognition, pp. 26–36 (2016). https://doi.org/10.1007/978-3-319-45886-1_3 . Springer
73. Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Yumer, E., Yang, M.-H.: Learning blind video temporal consistency. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 170–185 (2018). https://doi.org/10.1007/978-3-030-01267-0_11
74. Mallya, A., Wang, T.-C., Sapra, K., Liu, M.-Y.: World-consistent video-to-video synthesis. In: Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII, pp. 359–378. Springer, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58598-3_22
75. Rivoir, D., Pfeiffer, M., Docea, R., Kolbinger, F., Riediger, C., Weitz, J., Speidel, S.: Long-term temporally consistent unpaired video translation from simulated surgical 3d data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3343–3353 (2021). https://doi.org/10.1109/ICCV48922.2021.00333
76. Zhuo, L., Wang, G., Li, S., Wu, W., Liu, Z.: Fast-vid2vid: Spatial-temporal compression for video-to-video synthesis. In: European Conference on Computer Vision, pp. 289–305 (2022). https://doi.org/10.1007/978-3-031-19784-0_17 . Springer
77. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., *et al.*: Benchmarking 6dof outdoor visual localization in changing conditions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8601–8610 (2018). https://doi.org/10.1109/CVPR.2018.00897
78. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017). https://doi.org/10.1177/0278364916679498
79. Sakaridis, C., Wang, H., Li, K., Zurbrugg, R., Jadon, A., Abbeloos, W., Reino, D.O., Van Gool, L., Dai, D.: Acdc: the adverse conditions dataset with correspondences for robust semantic driving scene perception. IEEE Trans. Pattern Anal. Mach. Intell. , 48(3), 2970–2988 (2026). https://doi.org/10.1109/TPAMI.2025.3633063
80. Alam, M.S., Bazilinska, O., Bazilinskyy, P.: A global dataset of continuous urban dashcam driving. arXiv:2604.01044v2 (2026) https://doi.org/10.48550/arXiv.2604.01044
81. Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset. Computer vision and image understanding 178, pp. 30–42. (2019). https://doi.org/10.1016/j.cviu.2018.10.010
82. Wang, W., Yang, H., Li, B., Sun, J., Zhao, X., Xu, Z., Guo, Q., Min, H., Zhang, T., Yu, H.: Darkdriving: A real-world day and night aligned dataset for autonomous driving in the dark

environment. arXiv preprint arXiv:2603.18067 (2026) https://doi.org/10.48550/arXiv.2603.18067
83. Tan, X., Xu, K., Cao, Y., Zhang, Y., Ma, L., Lau, R.W.H.: Night-time scene parsing with a large real dataset. Trans. Img. Proc **30**, 9085–9098 (2021). https://doi.org/10.1109/TIP.2021.3122004
84. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1808–1817 (2015). https://doi.org/10.1109/CVPR.2015.7298790
85. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 6629–6640. Curran Associates Inc., Red Hook, NY, USA (2017). https://doi.org/10.5555/3295222.3295408
86. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a completely blind image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2013). https://doi.org/10.1109/LSP.2012.2227726
87. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. **21**(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050
88. Kwak, J.-g., Jin, Y., Li, Y., Yoon, D., Kim, D., Ko, H.: Adverse weather image translation with asymmetric and uncertainty-aware gan. arXiv preprint arXiv:2112.04283 (2021) https://doi.org/10.48550/arXiv.2112.04283
89. Li, X., Guo, X.: Spn2d-gan: Semantic prior based night-to-day image-to-image translation. Trans. Multi **25**, 7621–7634 (2023). https://doi.org/10.1109/TMM.2022.3224329
90. Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Larochelle, H.: Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). J. Mach. Learn. Res. **22**(164), 1-20 (2021). http://jmlr.org/papers/v22/20-303.html https://doi.org/10.48550/arXiv.2003.12206
91. Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control (MAQC) Society Board of Directors, Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C.S., Broderick, T., Hoffman, M.M., Leek, J.T., Korthauer, K., Huber, W., Brazma, A., Pineau, J., Tibshirani, R., Hastie, T., Ioannidis, J.P.A., Quackenbush, J., Aerts, H.J.W.L.: Transparency and reproducibility in artificial intelligence. *Nature***586**(7829), E14–E16 (2020) https://doi.org/10.1038/s41586-020-2766-y
92. Gundersen, O.E., Coakley, K., Kirkpatrick, C., Gil, Y.: Sources of irreproducibility in machine learning: A review. arXiv preprint arXiv:2204.07610 (2022). https://doi.org/10.48550/arXiv.2204.07610
93. Zheng, Z., Wu, Y., Han, X., Shi, J.: Forkgan: Seeing into the rainy night. In: European Conference on Computer Vision, pp. 155–170 (2020). https://doi.org/10.1007/978-3-030-58580-8_10 . Springer
94. Pizzati, F., Cerri, P., De Charette, R.: Comogan: continuous model-guided image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14288–14298 (2021). https://doi.org/10.1109/CVPR46437.2021.01406
95. Pizzati, F., Lalonde, J.-F., Charette, R.: Manifest: Manifold deformation for few-shot image translation. In: European Conference on Computer Vision, pp. 440–456 (2022). https://doi.org/10.1007/978-3-031-19790-1_27 . Springer
96. Li, X., Guo, X., Zhang, J.: N2d-gan: A night-to-day image-to-image translator. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2022). https://doi.org/10.1109/ICME52920.2022.9859906 . IEEE
97. Kennerley, M., Wang, J.-G., Veeravalli, B., Tan, R.T.: 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11484–11493 (2023). https://doi.org/10.1109/CVPR52729.2023.01105
98. Ning, J., Gong, M.: Enhancing night-to-day image translation with semantic prior and reference image guidance. In: Australasian Database Conference, pp. 164–182 (2023). https://doi.org/10.1007/978-3-031-47843-7_12 . Springer
99. Parmar, G., Park, T., Narasimhan, S., Zhu, J.-Y.: One-step image translation with text-to-image models. arXiv preprint arXiv:2403.12036 (2024) https://doi.org/10.48550/arXiv.2403.12036
100. Che, Z., Li, G., Li, T., Jiang, B., Shi, X., Zhang, X., Lu, Y., Wu, G., Liu, Y., Ye, J.: D^2-City: A large-scale dashcam video dataset of diverse traffic scenarios. arXiv preprint arXiv:1904.01975 (2019) https://doi.org/10.48550/arXiv.1904.01975
101. Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., Milford, M.: Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. Robotics: Science and Systems XI , 1–10 (2015). https://doi.org/10.15607/RSS.2015.XI.022
102. Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., Civera, J.: Mapillary street-level sequences: A dataset for lifelong place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2626–2635 (2020). https://doi.org/10.1109/CVPR42600.2020.00270
103. Li, B., Fu, C., Ding, F., Ye, J., Lin, F.: All-day object tracking for unmanned aerial vehicle. IEEE Trans. Mob. Comput. **22**(8), 4515–4529 (2022). https://doi.org/10.1109/TMC.2022.3162892