GDAŃSK UNIVERSITY OF TECHNOLOGY

The author of the doctoral dissertation: Efkleidis Katsaros
Scientific discipline: Information and Communication Technology

DOCTORAL DISSERTATION

Title of doctoral dissertation: Deep Video Multi-task Learning Towards Generalized Visual Scene Enhancement and Understanding

Title of doctoral dissertation (in Polish): Głębokie uczenie sieci wielozadaniowych dla zadań poprawy jakości i rozumienia wideo

Supervisor prof. dr hab. inż. Jacek Rumiński

signature

Gdańsk, year 2024

Summary

The rise of deep learning has advanced the field of artificial intelligence with numerous applications including but not limited to vision, language and game playing. In real-life scenarios, using deep learning to address just one vision task is inadequate for numerous industries and applications. Multi-task learning is a machine learning paradigm where a single model is trained to perform multiple related tasks simultaneously and bears some resemblance to the human brain's ability to process and learn various tasks in parallel. Much like how humans can transfer knowledge and skills learned in one domain to another, multi-task learning allows a model to leverage information learned from one task to improve performance on related tasks.

Until now, the scope of multi-task learning has been constrained to very specific visual tasks. Existing research predominantly centers on visual scene understanding tasks, with typically correlated task outputs. Consequently, both datasets and architectures have primarily revolved around those specific tasks. While numerous datasets have been proposed, they often lack labels for visual scene enhancement tasks. Additionally, previous studies have been primarily focused on static images, neglecting the valuable video data available in neighboring frames. Furthermore, multi-task architectures tend to rely on manual design choices driven by practitioner or researcher assumptions. Lastly, training these multi-task networks typically demands computationally intensive optimization methods, often yielding marginal benefits.

The goal of this thesis was to *develop efficient video multi-task convolutional architectures for a range of diverse vision tasks, on RGB scenes, leveraging i) task relationships and ii) motion information* to improve multi-task performance. The approach we take starts from the integration of diverse tasks within video multi-task learning networks. We present the first two datasets of their kind in the existing literature, featuring frame-level annotations for both visual scene enhancement and understanding. This thesis proposes novel architectures, capable of accommodating multiple tasks across various hierarchy levels. The second contribution of this thesis extends those findings into the MOST (Multi-Output, -Scale, -Task) model which exploits the inherent multi-scale nature of convolutional networks in a manner that benefits video multi-tasking. Thereafter, we propose a principled pruning approach inspired by NAS (Neural Architecture Search), named NSS (Neural Struc-

ture Search). NSS discovers a more effective MOST network, which boosts performance while simultaneously reducing computational requirements and parameter count. Lastly, we introduce ATB (Adaptive Task Balancing), an efficient training method that ensures tasks are trained at consistent rates with almost no additional computational cost, enabling a more balanced multi-task training process.

The contributions of this thesis are experimentally verified and part of the experimental results supporting this thesis has been published in several scientific papers. Practically speaking, the results in this thesis can help practitioners, interested in efficiently solving multiple tasks, by suggesting topological designs, architectural components and multi-task training regimes.

Streszczenie

Rozwój głębokiego uczenia się poszerzył dziedzinę sztucznej inteligencji o liczne zastosowania, w tym między innymi widzenie i język. W rzeczywistych scenariuszach wykorzystanie głębokiego uczenia się do rozwiązania tylko jednego zadania związanego z wizją jest odpowiednie dla wielu branż i zastosowań. Uczenie wielozadaniowe to paradygmat uczenia maszynowego, w którym pojedynczy model jest szkolony do jednoczesnego wykonywania wielu powiązanych zadań i jest w pewnym stopniu podobny do ludzkiego mózgu pod względem równoległego przetwarzania i uczenia się różnych zadań. Podobnie jak ludzie mogą wykorzystywać wiedzę i umiejętności nabyte w jednej dziedzinie do radzenia sobie z innymi problemami, uczenie wielozadaniowe umożliwia modelowi wykorzystanie informacji zdobytych podczas wykonywania jednego zadania w celu poprawy skuteczności w przypadku powiązanych zadań.

Do tej pory zakres uczenia wielozadaniowego ograniczał się do konkretnych kombinacji zadań. Istniejące badania skupiają się głównie na zadaniach związanych ze zrozumieniem scen wizualnych, których wyniki są zazwyczaj skorelowane. W rezultacie zarówno zbiory danych, jak i proponowane architektury sieci neuronowych skupiały się głównie wokół tych konkretnych zadań. Chociaż zaproponowano wiele zbiorów danych, często brakuje im etykiet dla zadań poprawy jakości obrazu. Ponadto poprzednie badania koncentrowały się głównie na obrazach statycznych, zaniedbując cenne dane z sąsiednich klatek dostępne w przypadku wideo. Co więcej, architektury wielozadaniowe zwykle opierają się na intuicyjnych wyborach opartych na założeniach inżynierów lub badaczy. Wreszcie, szkolenie sieci wielozadaniowych zazwyczaj wymaga metod optymalizacji wymagających intensywnych obliczeń, często przynoszących niewielkie korzyści.

W tej pracy badamy integrację różnorodnych zadań w ramach wielozadaniowych sieci przetwarzających wideo. Przedstawiamy dwa nowatorskie zbiory danych zawierające adnotacje na poziomie ramki. W naszej pracy wprowadzamy innowacyjne architektury, zdolne do obsługi wielu zadań na różnych poziomach hierarchii i rozszerzamy te ustalenia na model MOST (Multi-Output, -Scale, -Task). Następnie proponujemy metodę redukcji modeli inspirowaną NAS (Neural Architecture Search), nazwaną NSS (Neural Scale Search). NSS odkrywa bardziej efektywną sieć MOST, która zwiększa wydajność, jednocześnie zmniejszając wymagania obliczeniowe i liczbę parametrów. Na koniec przedstawiamy ATB (Adaptive

Task Balancing), wydajną metodę trenowania wielozadaniowych sieci neuronowych, która zapewnia uczenie zadań ze stałą szybkością, niemal bez dodatkowych kosztów obliczeniowych, umożliwiając bardziej zrównoważony proces treningu.

Wkład tej tezy został zweryfikowany eksperymentalnie, a część wyników eksperymentów potwierdzających tę tezę została opublikowana w kilku artykułach naukowych. W praktyce wyniki tej pracy mogą pomóc praktykom zainteresowanym skutecznym rozwiązywaniem wielu zadań, sugerując projekty topologiczne, komponenty architektoniczne i procedury uczenia wielozadaniowego.

Acknowledgements

I would like to sincerely thank my supervisor, Professor Jacek Ruminski, for his supervision, advice and freedom granted to successfully pursue my studies. My sincere thanks go to all my colleagues and supervisors for their assistance and collective efforts towards our shared goals. Lastly, but certainly not least, I wish to express my immense gratitude to my friends and family in Greece, Poland and the Netherlands. Their belief in me and their support have been a constant source of motivation and strength throughout my doctoral journey.

List of abbreviations

The following abbreviations has been used throughout this thesis.

ML	Machine Learning
DL	Deep Learning
STL	Single-task Learning
MTL	Multi-task Learning
MTO	Multi-task Optimization
MO	Multi-output
MS	Multi-scale
MOST	Multi-output, -scale, -task
NAS	Neural Architecture Search
NSS	Neural Scale Search
ATB	Adaptive Task Balancing
LS	Linear Scalarization
DWA	Dynamic Weight Averaging
RWL	Random Weight Loss
$\mathcal{O}_i^s$	Output at scale s^* for task i
$\mathcal{L}_i^s$	Loss function for scale s^* for task i
$\mathcal{D}$	Dataset
$\mathcal{C}_{c,k\times k}$	convolution with c output channels & $k \times k$ kernel
$\mathcal{C}^D_{c,k\times k}$	deformable convolution with c output channels & $k \times k$ kernel
$\mathcal{RB}_c$	residual block with c output channels
$\mathcal{RDB}_{c,g}$	residual dense block with c output channels & growth g
$\mathcal{RDM}_{c,g}$	residual dense module with c output channels & growth g
$\mathcal{RDM}_{\mathcal{MS}c,g}$	multiscale residual dense module with c output channels & growth g
$\mathcal{E}$	Multi-task encoder
$\mathcal{D}^s_{task}$	Task-specific decoder at scale s
GAP	Global Average Pooling
MP	Max Pooling
ECA_k	Efficient Channel Attention with $1 \times k$ kernel
$MECA_k$	Modulated Efficient Channel Attention with $1 \times k$ kernel
B_t^s	Input corrupted frame at time step t and scale s^*
R_t^s	Restored output at time step t and scale s^*
M_t^s	Segmentation output at time step t and scale s^*

H_t^s	Homography output at time step t and scale s^*
f_t^s	MTL features describing a (stack of) frame(s) at time step t and scale s^*
$f_{a,t}^s$	Aligned MTL features describing a (stack of) frame(s) at time step t and scale s^*
$W_H(f_{t-p})^s$	Homography-aligned MTL features describing a frame at time step $t-1$ and scale s^*, aligned to frame t
$W_{OF}(f_{t-p})^s$	Flow-aligned features MTL describing a frame at time step $t-1$ and scale s^*, aligned to frame t
F_t^s	Attended multi-task features describing a (stack of) frame(s) at time step t and scale s^*
g_t^s	Shared decoder features at time step t and scale s^*
G_t^s	Recurrent multi-task features at time step t and scale s^*

∗ Wherever the superscript s is missing throughout this dissertation, the highest scale is implied, i.e. $s = 1$, but is ommitted for notational brevity.

Contents

Chapter 1

Introduction

1.1 Foreword and Motivation

Recent years have seen tremendous progress in the application of machine learning models to the real world. Biologically inspired neural networks, led to significant improvements in various applications, including image recognition, natural language processing, and speech recognition. Deep learning models achieved near human-level performance in tasks such as image classification [1, 2, 3] and object detection [4, 5] with applications in fields like medical imaging and autonomous driving. Progress in Natural Language Processing led to the development of pre-trained language models like BERT [6], GPT-3 [7], and GPT-4 [8], which showcased remarkable language understanding and generation capabilities. Reinforcement learning algorithms achieved impressive results in playing complex games like Go [9], chess [10], and video games [11], outplaying human experts. Generative models [12] made it possible to produce realistic images, videos, and audio. They have applications in art, entertainment, and content creation. More recently, researchers tackled the protein folding problem [13] with unprecedented accuracy, bypassing the efforts of a whole research community and accelerating research in nearly every field of biology.

The emergence of AlexNet [2] in 2012 significantly accelerated the deep learning revolution in AI. This pioneering convolutional neural network (CNN) achieved a groundbreaking leap in image recognition accuracy during the ImageNet Large Scale Visual Recognition Challenge [14], fundamentally reshaping the landscape of computer vision. While CNNs had already existed for decades [15], their full potential remained unknown until AlexNet. The authors utilized multiple subsequent layers of convolutions, non-linear activations, pooling operations and dropouts, trained with heavy data augmentation to reduce overfitting and enable the optimization of a non-convex, highly parameterized, and highly non-linear function. What made it feasible to train such a computationally intensive model, was AlexNet's utilization of GPUs to train the network, setting a milestone in the field. This was then made possible with the utilization of the Caffe framework [16], setting the stage for the subsequent rise of TensorFlow [17] and PyTorch [18], which have since become

synonyms of the latest deep learning advancements.

Convolutional neural networks (CNNs) have achieved remarkable success since then, in two crucial areas of computer vision, visual scene enhancement and visual scene understanding, with huge industrial impact. In the domain of scene enhancement, CNNs benefit numerous sectors where image quality and precision are important. In healthcare, for instance, CNNs assist radiologists by enhancing medical images or by making diagnoses more accurate. In the satellite imaging industry, they improve the clarity of remote sensing data, with applications in environmental monitoring and disaster response. Even the entertainment industry benefits from scene enhancement, where CNNs are used to upscale low-resolution video content, providing viewers with a more immersive and enjoyable experience. Similarly, scene understanding has brought automation and efficiency to numerous industries. In manufacturing, CNN-based quality control systems can rapidly inspect and identify defects on production lines, ensuring consistency and reducing errors. The retail sector uses CNNs for shelf monitoring and inventory management, optimizing stock levels and improving customer experiences. Furthermore, CNNs are deployed on unmanned aerial vehicles in different industries, to reach hardly accessible areas and inspect e.g. wind turbines for damage, or detect fire in forests. In both scene enhancement and scene understanding, CNNs enhance visual data for human interpretation and automate processes across industries, ultimately driving innovation and efficiency.

In many fields, visual scene enhancement and understanding are simultaneously needed. In medical imaging, assisting doctors for minimally invasive surgeries requires the improvement of the image quality and segmenting the regions of interest in operations such as colonoscopies and intra-oral interventions. Environmental monitoring systems, such as satellite imagery or drone footage, use scene enhancement to reduce blur or haze incurred by flying instabilities or the atmosphere respectively. Simultaneously, visual scene understanding algorithms can be applied to detect changes in landscapes, identify forest fires, assess crop health, or monitor pollution levels. Interestingly, scene enhancement is capable of complementing understanding by critically improving its performance, e.g. in road scenes, illustrating a use case for autonomous driving. What is more, tasks like pedestrian detection, lane recognition, and object tracking, which enable self-driving cars to navigate safely, are easier in enhanced rather than low-light videos.

To address those issues, one could use multiple, individual single-task CNNs for each specific task, however it would be rather limiting to do so. Unlike the human brain, which models correlations between different functionalities of the prefrontal cortex, single-task approaches overlook the interrelationships between tasks. This leads to reduced performance and inefficient utilization of model capacity, as expressed by its parameters. Additionally, the training and deployment of per-task neural networks is resource intensive. Handling multiple tasks necessitates multiple forward passes through distinct networks, even when a foundational understanding of image features is shared among these tasks.

Another limitation of multi-task state-of-the-art CNNs is that they rely on static image inputs, even when dealing with video streams. In such cases, individual frames are often processed independently, which simplifies the analysis but is not an optimal for the other tasks. Firstly, leveraging spatiotemporal information can improve performance in both scene enhancement and understanding tasks. Secondly, acquiring the ability to exploit information from neighboring frames enhances the model's adaptability to diverse data sources. Motion allows the networks to learn how to reason more effectively about the scene spatial layout. Last, the capacity to track object pixels across frames results in more robust features, leading to improved image enhancement and recognition accuracy.

1.2 Scope and Contributions

The goal of this thesis is to develop parameter- and runtime-efficient convolutional neural networks for video multi-task learning that are capable of exploiting the spatiotemporal domain to learn better frame representations, for RGB video scenes, and the multi-task interactions to effectively accommodate diverse tasks of different hierarchies, encompassing both visual scene enhancement and understanding.

Multi-task learning has attracted significant research interest. It effectuates synergic network topologies to increase performance among tasks and accelerates inference by alleviating the necessity for multiple forward passes over dedicated single-taskers. Many works attempted to leverage synergic information across tasks to improve performance. MTAN [19] learnt a global multi-task feature pool and employed soft attention mechanisms to query the pool and retrieve the features mostly relevant to each of the task-specific decoders. PAD-Net [20] yielded a set of initial multi-task predictions and refined them with an attention-guided message passing mechanism for distillation. ATRC [21] enabled multi-task cross-talk by learning different types of attention mechanisms for different tasks. MTI-Net [22] propagated and distilled features and outputs to the decoders across tasks.

The aforementioned approaches are not suitable for quality enhancement because they are specifically tailored for scene understanding datasets and tasks like segmentation, depth estimation, and surface normal estimation (e.g., [23, 24, 25]). These tasks operate within a similar leve hierarchy, where information exchangeability is profound. In Fig. 1.1, we visualize the ground truth maps for semantic segmentation and depth estimation from the Cityscapes [23] and NUYv2 [25] multi-task labelled datasets. On Cityscapes, depth maps are associated with segmentations via the contours. On NUYv2 segmentation maps can assist in estimating the depth values. In both scenarios, the solutions for each task can straightforwardly complement and improve the performance of the other. A similar example is illustrated in Fig. 1.2, where we visualize the results of UberNet [26], this time using the PascalVOC dataset [27]. Another limitation of current multi-task networks is that they

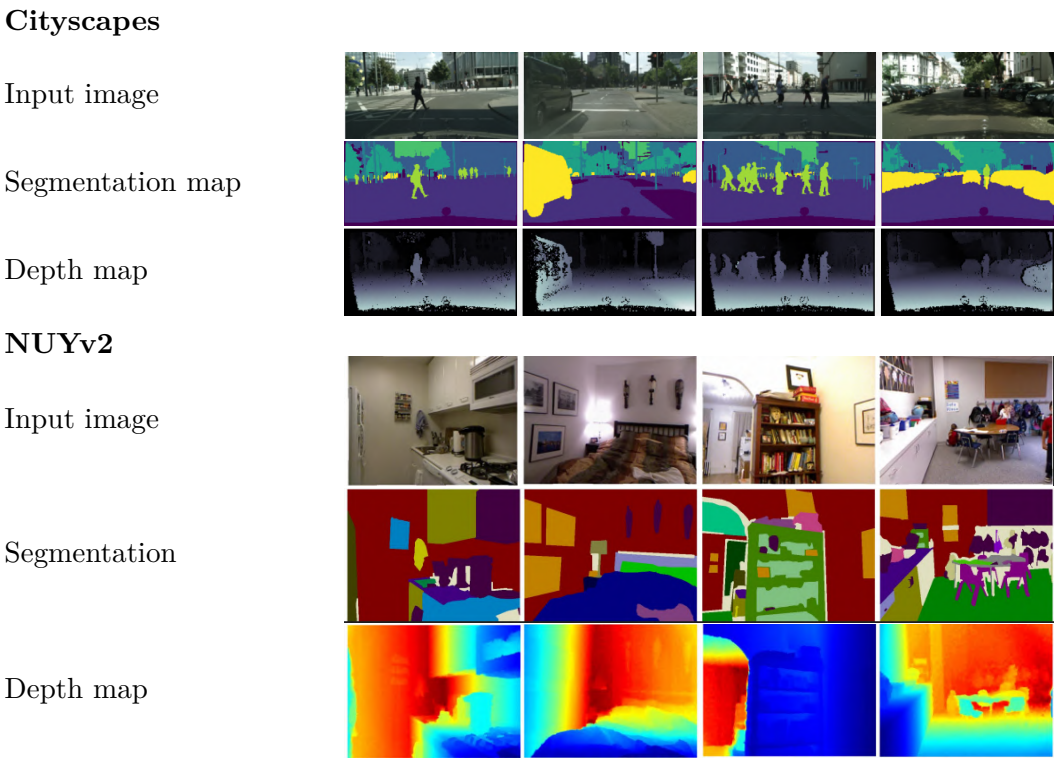


Figure 1.1: Ground truth labels for semantic segmentation and depth estimation on the Cityscapes (top) and NUYv2 (bottom) datasets. On Cityscapes, depth maps are associated with segmentations via the contours. On NUYv2 segmentation maps can assist in estimating the depth values. In both cases, each task solution can facilitate the other.

make dense task predictions in static images, failing to exploit the potential benefits of information aggregation across successive video frames within the temporal domain.

The integration of diverse tasks into practical video applications, such as deblurring and denoising or segmentation, particularly, when the degradation artifacts occur across the the whole image, remains largely unexplored in the existing literature. Initially, considering that deblurring and denoising are foundational tasks for visual scene enhancement and occur in many settings together due to known noise-vs-blur issue [28], we study them together. Notably, this doctoral thesis is the first to consider deep multi-task learning for those tasks. Their combination is challenging, since they involve relatively uncorrelated challenges – knowledge of one task does not significantly assist the other. We propose a multi-task architecture to efficiently address both these tasks. Subsequently, we extend our investigation to include one visual scene enhancement task alongside a scene understanding task, merging video deblurring and object segmentation. These inquiries give rise to our primary, two-fold hypothesis:

First hypothesis: Visual scene enhancement tasks such video denoising and deblurring can be effectively integrated through a lightweight, deep multi-task network at improved performance and computation compared to existing single-task approaches. Similarly, we assert that object segmentation is capable of complementing video deblurring in a synergic architecture. Our hypothesis is limited in the RGB domain, yet this thesis progresses to explore video footage, instead of images, since both scene enhancement and understanding tasks share benefits from modelling inter-frame motion.

Motion information is naturally inherent in videos. It enables frame alignment, which in turn facilitates the concurrent tasks. In the subsequent hypotheses explored in this dissertation, we further consider two more diverse tasks into our task pool, homography estimation and low-light enhancement, as in color mapping. Homography estimation serves as a proxy of the camera motion, while color mapping learns brighter, more vivid colors for video scenes. However, the increasing diversity and number of tasks introduce complexities onto the architectural design. We seek answers to questions such as how multi-task architectures perform when dealing with a broader and more varied array of tasks, how to allow for task interactions for better multi-task learning, and what constitutes an optimal architectural design for this purpose, comparing it against state-of-the-art single-task convolutional networks.

Second hypothesis: We posit that a multi-task network topology exists, capable of exploiting the multi-scale nature of convolutional networks to accommodate visual scene both enhancement and understanding tasks while allowing them to communicate. Such a network would scale up to a multitude of both low- and higher-level tasks. We further assume that such a deep video multi-task network, should be at minimum,

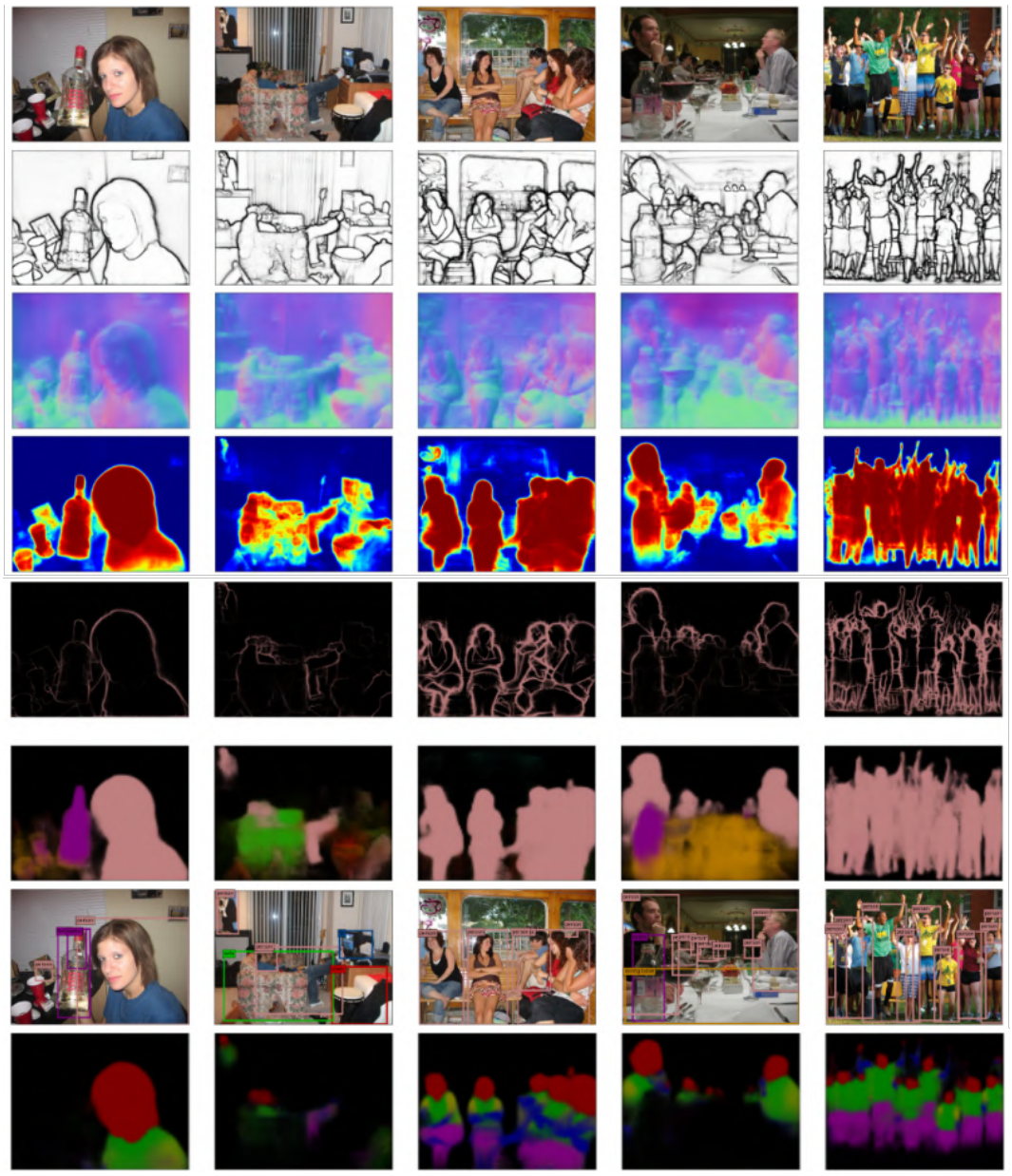


Figure 1.2: UberNet [26] outputs on the PascalVOC dataset. Inter-task affinities are high, and each task solution, again, facilitates the other.

competitive with state-of-the-art systems of networks, at lower compute.

Throughout this doctoral study, two dental video datasets, with multi-task labels, are recorded and processed. The first dataset is recorded in lab conditions and is publicly available [29]. The second dataset consists of real intra-oral surgeries, performed by dental experts, and will also be made publicly available with an upcoming journal publication.

Even with a topologically appropriate architecture, effective allocation of multi-task resources is not guaranteed. Some tasks might require different parameter capacity at different parts of the architecture. Prior research [30] has demonstrated significant performance enhancements in single-output, multi-task architectures. In a similar line of thought, this thesis argues that not all scales are beneficial for all tasks. The visual scene enhancement tasks operate better at higher scales while scene understanding is typically performed at lower scales to benefit from larger effective receptive fields. Instead of making rigid assumptions on the impact of the scales at the tasks and their interactions, this thesis casts the architectural scales-to-tasks structure into an optimization problem. It assumes it can discovered from the data itself with simple back-propagation, in a NAS inspired approach [31].

Third hypothesis: We hypothesize that the multiple scales of a multi-task convolutional network can be allocated to tasks, such that the per-task intricacies and the across-tasks interactions are optimally accommodated. Rather than manually designing such an architecture with rigid assumptions on the impact of each scale on each task, we posit that there are benefits to allowing the network to learn these relationships in an end-to-end manner.

Despite their efficiency and – under conditions – improved performance, training multi-task networks involves the minimization of objectives exhibiting different magnitudes (norms) or directions (angles) and training speeds resulting in the dominance of some tasks over some others. Multiple multi-task optimization works (MTOs) have been proposed to alleviate the issue of differences in the norms or the angles of the multi-task gradients. Learnable balancing methods [32, 33, 34] involve learning parameters to balance the gradients or the loss weights while adaptive balancing methods [19, 35] employ heuristics. Most MTOs, however, require access to the per-task gradients to synthesize an enhanced weight update, but one major drawback is that memory and, thus, required time grows linearly with the number of tasks. This issue is shown in all [33, 36, 37]. Recent research [38, 39, 36] casts doubt on the effectiveness of MTOs. [38] show that MTO methods fail to outperform traditional approaches despite the added complexity and computational demands. All aforementioned methods with the exception of [19], attempt to address the different magnitudes and/or directions of gradients but neglect the diverse multi-task training speeds. In contrast, this thesis addresses explicitly the latter, with a a focus on training efficiency to facilitate its applicability on practical scenarios:

Fourth hypothesis: Tasks may exhibit varying training speeds, im-

plying that the advancements achieved in task A during training may significantly differ from those in task B. Efficiently aligning the training progress of these tasks can yield more stable multi-task learning.

1.3 Thesis outline

The rest of this thesis is structured as follows. Chapter 2 provides essential background information and explores the relevant prior research. It discusses single-task learning architectures for a range of diverse, visual scene enhancement and understanding tasks. Subsequently, it reviews multi-task architectures for different tasks and discusses their limitations. Last, we discuss briefly multi-task optimization challenges and the public datasets. The Chapter concludes with some literature gaps this thesis attempts to answer.

Chapter 3 shows that solving two visual enhancement tasks i.e. denoising and deblurring, in a unified, cascaded manner is better than solving them separately. Akin to the physics model behind the degradation factors, a novel multi-task, cascaded network is proposed, enriched with components that enhance the feature representations. We introduce deformable convolutions to align neighboring frames efficiently, investigating the trade-offs between blur and various levels of noise. We evaluate our architecture using a publicly available outdoor dataset and demonstrate its applicability across different scenarios. An ablation study further investigates on the proposed architecture and showcases that with minor performance drop, it is even capable of handling both visual enhancement tasks in one go, i.e. without allocating separate decoders for each task. Chapter 3 essentially confirms the first part of our first hypothesis.

Moving on to Chapter 4, we merge video deblurring and segmentation. While the former attempts to enhance the captured scene's quality, the latter is a scene understanding task that answers where something is. Each of the tasks bears distinct challenges, yet information from the one might be complementary to the other. Moreover, visual cues from previous frames can facilitate their multi-tasking even further. In this Chapter, we employ a dynamic kernel approach to leverage consecutive frame information, showcasing improved deblurring performance compared to state-of-the-art, while achieving very good segmentation outputs. Given dental video footage, we show that segmentation of the dental tooltip, in this scenario, assists deblurring by improving its performance. Chapter 4 revolves around the second part of the first hypothesis of this work.

In Chapter 5, we integrate video deblurring, denoising, color mapping, segmentation, and homography estimation, introducing the first dataset with labels for all these tasks. Notably, motion estimation through homographies is utilized for frame alignment. We propose a multi-scale, multi-output, multi-task architecture that propagates the outputs bottoms-up, from the lower to the higher image scale, enabling task interaction and output refinement. Moreover, our approach provides

insights on the impact each scale on each task's performance. We demonstrate the synergic capabilities of our general, MOST architecture and showcase that it outperforms multiple single-task networks on the newly-proposed dataset. MOST affirms the second hypothesis of this work.

In Chapter 6, we combine MOST with Neural Architecture Search into a framework dubbed Neural Scale Search, to learn the optimal resource allocation. More specifically, we argue that not all scales are beneficial for all tasks. The visual scene enhancement tasks operate better at higher scales while scene understanding is typically performed at lower scales to benefit from larger effective receptive fields. NSS discovers a more efficient and high-performing architecture compared to the general MOST. Additionally, we showcase the effectiveness of Adaptive Task Balancing over state-of-the-art, relevant optimization methods, achieving superior performance and quicker convergence with enhanced training stability. The same Chapter discusses the second dataset related to this dissertation. Consequently, it validates the two last theses of this work.

Finally, Chapter 7 discusses the contributions reported in this thesis, which provide discussion and directions for future work.

Chapter 2

Background and related work

Deep learning has revolutionized the field of artificial intelligence by enabling the creation of highly sophisticated models capable of learning intricate patterns and representations directly from data. Within deep learning, several subcategories exist, each tailored to different learning paradigms. In supervised learning [40], models are trained on labeled datasets, learning to map input data to corresponding output labels. Unsupervised learning [41], on the other hand, involves learning patterns and representations from unlabeled data alone, without explicit supervision. Semi-supervised [42] learning utilizes both labeled and unlabeled data, leveraging the abundance of unlabeled data often available as regularization. Deep reinforcement learning [43] focuses on training agents to make sequential decisions through interaction with an environment, maximizing cumulative rewards.

Generally, supervised learning tends to be highly effective when labeled data is abundant and the task is well-defined, allowing models to learn directly from examples with clear feedback. Reinforcement learning is powerful for tasks involving sequential decision-making and dynamic environments, where agents must learn optimal strategies through trial and error. Semi- or unsupervised learning are valuable for tasks where labeled data is scarce or unavailable, enabling models to discover underlying structures and representations only from the data distributions. When the training data are diverse enough and their labels are not noisy, supervised learning yields higher performance compared to learning from some lesser supervision paradigm for single tasks. Following the same narrative, this thesis studies multi-task learning in a supervised setup, where all labels are provided for all tasks.

In the following sections, we discuss some characteristic and well-studied computer vision tasks, for both visual scene enhancement and understanding. Specifically, in Section 2.1 we review supervised, single-task approaches, for the tasks addressed in this thesis, i.e. video deblurring, denoising, low-light enhancement or color mapping, semantic segmentation and homography estimation. Moving on to Section 2.2, we discuss supervised, multi-task works, relevant to the aforementioned tasks. In line with the scope of this thesis, we limit the reviewed literature to convolutional architectures and do not consider transformers [44, 45]. As we hypothesize,

that the nature of tasks typically addressed in multi-task learning is associated with the availability of data testbeds, Section 2.3 sheds light on the publicly available datasets. Finally, Section 2.4 lists the primary literature gaps detected, that this thesis aims to fill in.

2.1 Supervised learning

In this dissertation, we consider diverse tasks categorized into two overarching domains: visual scene enhancement and visual scene understanding. Visual scene enhancement tasks aim at improving the quality and appearance of images. At the low level, basic image processing techniques such as denoising or contrast enhancement are employed to enhance image clarity and visual appeal. Moving to mid-level tasks, efforts focus on enhancing specific aspects of scenes, such as edges or textures for deblurring, to make certain features more aesthetically pleasing. High-level tasks in visual scene enhancement include more complex operations, such as super-resolution, to upscale the image resolution and detail, or image inpainting to fill in missing parts and improve overall appearance. In subsection 2.1.1, we revise important video deblurring, denoising and color mapping literature.

On the other hand, visual scene understanding tasks revolve around extracting meaning and context from images. Low-level tasks involve fundamental operations like optical flow, homography and depth estimation, providing the groundwork for deeper analysis. Mid-level tasks delve into the scene spatial layouts, aiming to understand where is what, such as in object detection and image segmentation. Finally, high-level tasks in visual scene understanding include more sophisticated operations such as scene classification, and image captioning. In subsection 2.1.2, we review multiple homography estimation and semantic segmentation works.

2.1.1 Visual Scene Enhancement

Visual scene enhancement, commonly referred to also as restoration, refers to a range of early vision tasks such as denoising [46, 47], deblurring [48, 49, 50] and low-light enhancement [51]. In the deep learning literature, those tasks typically addressed via encoder-decoder architectures. Recently, generic architectures were proposed [52, 53], capable of successfully addressing multiple restoration tasks with a single topology, optimized for task at hand. The underlying assumption is that the challenges are similar, i.e. fusing local i.e. object-based and temporal context, to learn some complex transformation of the input image. The first common characteristic between architectures in this category, is that they retain a high number of computations at higher feature resolutions (scales) to avoid the loss of spatial information that would degrade the output image. When temporal context exists, i.e. video frames are available, those architectures share a second characteristic; they attempt to align frames or features to borrow visual cues by learning through

implicit pixel trajectories. Below, we review various popular methods for the tasks of deblurring, denoising and low-light enhancement.

Deblurring

Image deblurring aims to remove the blur from an input image, typically without any knowledge on its type or level. While traditional approaches estimate a blur kernel, deep learning methods bypass the ill-posedness of estimating the blur kernel, by relying on the expressiveness of deep convolutional networks. Nah et al. [54] deblur the input images at multiple scales in a coarse-to-fine manner, utilizing three convolutional networks, each with several residual blocks [55] at a fixed scale. Tao et al. [56] deblur the input images at multiple scales via a coarse-to-fine but recurrent design. Different to [54], each scale utilizes an encoder-decoder network and propagates both a memory state and the output image to the next one. Yuan et al. [57] employ deformable convolutions [58] to estimate the blur kernel by adaptively adjusting the kernel offsets to the input features, further supervised by optical flow. Suin et al. [59] achieve faster runtimes compared to previous methods by designing lightweight attention modules combined with adaptive kernels at three different multi-patch hierarchies. Zamir el al. [60] employ a three-stage network comprising of two multi-scale and one single-scale sub-networks where they allow for feature communication between different scales and stages of the proposed solution.

To deal with dynamic scenes and exploit temporal context more effectively, video deblurring methods further attempt to align the previous frames with the current one. Su et al. [61] propose an encoder-decoder architecture to align the input frames via its intrinsic multi-scale property and show that warping the input frames with optical flow introduces negligible performance gains but significant warping artifacts. In similar research lines, Zhou et al. [50] perform implicit frame alignment on the feature level by learning alignment kernels to leverage spatiotemporal features and overcome inaccurate optical flow estimation. Wang et al. [62] propose EDVR, a general purpose restoration network for multiple visual scene enhancement tasks including deblurring, denoising and super-resolution. The authors perform feature-level alignment with a multi-scale cascaded module of deformable convolutions. This work was pioneering since the authors addressed all tasks with the same network, trained each time for the task at hand. Zhong et al. [48] introduce a recurrent network that extracts the features frame-wisely and pre-processes them with spatio-temporal attention module that emphasizes on the important ones to be passed to the reconstruction decoder that generates the output image. Pan et al [63] propose again to deblur using optical flow, but this time they estimate it dynamically to reduce the artifacts. While their method comes with high performance, it is computationally intensive to both train and evaluate.

Denoising

Image denoising aims to reduce noise from a noisy image. Deep learning approaches bypass the necessity for hand-crafted features and the manual optimization of the various hyper-parameters with stacked convolutional layers followed by non-

linear activation functions [64]. Santhanam et al. proposed a Recursively Branched Deconvolutional Network (RBDN) [65], to learn a multi-context image representation using an efficient recursive branching scheme. This multi-context representation is then inputted further to a convolutional neural network comprising of a series of convolution filters at the same, or higher feature resolutions, to avoid loss of spatial information. Interestingly, convolutional neural networks have shown to be able to denoise different types and levels of noise with a single model. One of the early, pioneering works pointed that out [66]. Zhang et al. showcased that convolutional networks were able to perform denoising regardless of the level, with a single network. Their work further investigated on the design of the architectures and incorporated residual learning and batch normalization into their DnCNN network, to speed up training and increase performance. Following up their work, the same authors proposed FFDNet [67] to accelerate DnCNN while retaining or even improving performance.

Video denoising has similarly attracted vast research interest by the deep learning community. Here, one of the main challenges compared to image denoising, is to effectively employ information from neighboring frames to restore the current. Prior works rely on recurrent [68] and kernel-predicting [69] neural networks, however the results were far from the current state-of-the-art. In terms of convolutional neural networks, one of the first works was VNLnet [70], a non-local network that leverages convolutions along with a self similarity search technique. Their network searches for similar patches and their affinity is utilized to estimate the denoised output. Tassano et al. proposed DVDNet [71], a method which warps neighboring frames to the central one, and denoises it with two cascaded denoising stages. Despite the very good performance, the method relied on optical flow to warp the images and was impractical for real world application environments. In ViDeNN [72], the authors again denoised the central frame with a two-stage approach, however, their method does not compensate the motion, resulting in faster runtimes but lower performance compared to DVDNet. Following a similar narrative, Tassano et al. [46] kept the two-stage denoising structure, but employed a multi-scale, UNet-based [73] architecture to compensate the motion implicitly, achieving remarkably high efficiency and runtimes for the first time. In another work discussed earlier [62], Wang et al. proposed a general purpose restoration network which successfully handles video denoising, yet the introduced method is computationally intensive. Most of the aforementioned works are generally expensive to compute and rely on future frames, thus introducing latency on deployment environments. In a work supporting the findings of this doctoral thesis, Ostrowski et al. [47] proposed BP-EVD, a more performant algorithm that caches the feature computations, achieving a two-fold increase in FPS enabling real-time deployment for the first time on consumer grade GPUs.

Low-Light Enhancement

Supervised low-light enhancement methods require paired training data consisting of a low-light image and its corresponding well-lit version, known as the

ground truth. This allows the network to learn the relationship between the two and apply the necessary adjustments to enhance new low-light images. One example [74] involves a two-branch network that separates the image into illumination and reflectance maps. These maps are then used to train a separate UNet-based network to generate the final enhanced image. Another approach [75] focuses on generating realistic-looking dark images through specific transformations and noise addition. This data is then used to train a deep learning architecture with multiple subnetworks. The first subnetwork estimates the illumination, guiding the enhancement process for underexposed areas. The second subnetwork removes noise, while the third combines both tasks simultaneously. Finally, a fourth subnetwork refines the contrast to address limitations caused by the pixel-wise adjustments. Self-supervised learning methods, on the other hand, don't require paired training data. They operate under the assumption that low-light conditions can vary significantly across environments. These methods [76, 77] leverage deep learning architectures to analyze the low-light image itself and estimate an image-specific curve. This curve acts as a roadmap for adjusting the dynamic range of each pixel, ultimately enhancing the image quality without the need for a reference ground truth image. This approach offers greater versatility as it can be applied to various lighting conditions without requiring specifically paired data for each scenario. However, for applications where training data can be effectively synthesized, as with beam splitters, for instance, supervised learning is preferable.

2.1.2 Visual Scene Understanding

Visual scene understanding tasks, include a wide range such as the low-level optical flow or homography ones, for motion estimation and mid-level image segmentation or object detection, for understanding of the scene spatial layout. Motion estimation methods attempt to compute either the camera motion with a homography matrix or the pixel-wise motion with dense, optical flow maps. As they are low-lever tasks, they are crucial to a wide variety of downstream tasks such as image registration, stitching and video stabilization. Moreover, motion estimation is widely leveraged visual scene enhancement methods, implicitly or explicitly, as discussed in the previous subsection. Segmentation or object detection similarly allows for the identification of regions of interest to provide direct insights of the visual scene. Notably, both segmentation and motion estimation methods, require a coarse understanding of the image scene to reason i.e. in terms of salient objects. In this subsection, we discuss widely known models for homography estimation and semantic segmentation.

Homography Estimation

Recently, numerous deep homography estimation methods have been proposed with remarkable success. De Tone et al. [78] proposed a VGG-based [3] network to estimate the homography between two input images where the scene is assumed to be static. Furthermore, its authors adopted the Direct Linear Transform to reparam-

eterize the homography matrix into offsets and alleviate optimization issues related to different magnitudes of the homography target components. In another work, Chang et al. [79] employed the inverse compositional Lucas-Kanade [80] iterator on the convolutional features of two input images. They estimate and compensate the motion offsets iteratively, in a coarse-to-fine manner. Recently, Le et al. [81] proposed HMG, which adopted the idea behind PWC-Net [82] for optical flow estimation, to construct a cascade of networks operating at multiple scales, in order to estimate and refine the motion at each scale, in a residual fashion. Their work extended homography estimation to dynamic, i.e. scenes with independently moving objects. In a similar fashion, Cao et al. [83] extended the idea of RAFT [84] from optical flow to homography estimation. The authors followed RAFT in that they opted to refine the homography iteratively without performance divergence.

Semantic segmentation

Semantic segmentation has been enjoying immense progress since the rise of deep learning with numerous applications. Initially designed for medical image segmentation, UNet [73] was a pioneering work. Equipped with a symmetric pixel-to-pixel architecture, and enhanced with skip connections to circumvent the loss of spatial information, it was one of the first successful approaches in medical imaging and followed as the pre-requisite building block in a vast pool of both enhancement and understanding tasks. Several modifications have been proposed following its success, augmenting the UNet core idea with nested skip connections [85], dense connections [86], attention blocks [87] or its 3D counterpart [88]. One of its most advanced extensions, UNet++ [89] redesigned the skip pathways in the UNet architecture to enable better multi-scale information exchange and excelled at segmenting objects of different sizes. In 2018, Chen et al. introduced DeepLabV3+ (DLV3+) [90] another state-of-the-art segmentation network, highly performant on multiple datasets. DLV3+ entailed spatial pyramid pooling modules with dilated convolutions to encode multi-scale contextual information by probing the incoming features with filters or pooling operations at multiple rates. Its inherently augmented field-of-view and the fast runtimes renders DLv3+ the go-to architecture for generic segmentation tasks. Conversely, when smaller objects are to be segmented, the UNet variants are more effective.

2.2 Supervised multi-task learning

In this section, we review a series of largely influential supervised multi-task learning works. In visual scene enhancement, multi-task learning is still an emerging paradigm. Tasks including image super-resolution, denoising, deblurring, dehazing, and colorization are important in enhancing visual scene quality. Collectively addressing those tasks improves image clarity, reduces noise, restores details and enhances overall visual perception. Visual scene enhancement further assists other downstream vision tasks performed by computers. By improving inputs before feed-

ing them to computer vision algorithms, performance in detecting and recognizing objects increases [49] even under challenging conditions, such as low lighting, noisy or blurry environments [91]. In subsection 2.2.1, we review supervised, multi-task methods which consider multiple such visual scene enhancement tasks.

In the context of visual scene understanding, multi-task learning has been heavily studied in tasks such as object detection, semantic or instance segmentation, depth, saliency and surface normal estimation, and scene classification. These tasks collectively contribute to a more comprehensive understanding of visual data, enabling intelligent systems to interpret their environment more effectively. In subsection 2.2.2, we discuss supervised, multi-task methods for visual scene enhancement tasks. In subsection 2.2.3, we study a few multi-tasking methods addressing low image quality and scene understanding of some sort, simultaneously. Subsection 2.2.4 and 2.2.5 discuss and review the multi-task optimization challenge and the public datasets, while Subsection 2.2.6 closes with some points which we find less discussed in related works.

2.2.1 Visual Scene Enhancement

Only a few works attempted multi-task learning in visual scene enhancement. Cui et al. [92], for instance, addressed low-resolution and motion blur in face images simultaneously, with a multi-task approach that is based on two separate networks. The two networks were employed concurrently, to alleviate the accumulation of errors from one task to another. Furthermore, they were equipped with multi-scale feature fusion, channel and spatial attention modules to increase performance. However, the method only shared gradients and no parameters, as in soft-parameter sharing [93], being thus inefficient. Yu et al. [94] addressed the absence of multi-task learning methods in visual scene enhancement for medical image analysis and introduced an approach to simultaneously tackle super-resolution and denoising in medical images. The authors proposed a multi-task generator and integrated it with a discriminator, to learn better reconstruction features. This study however considered only MRI images.

In another line of research, a few works suggested to address multiple enhancement tasks with single-task encoder-decoder architectures. Zhou et. al [91] tackled the tasks of low-light enhancement and deblurring of real world videos. Despite addressing two visual scene enhancement tasks, the authors here did not employ a multi-task architecture but fused instead the two tasks into one dense prediction output, and relied on the internal, convolutional network filters to learn both at once. Xu et al. [95] presented a parametric representation called Deep Parametric 3D Filters (DP3DF), which incorporated local spatiotemporal information to enable simultaneous denoising, illumination enhancement, and super-resolution.

2.2.2 Visual Scene Understanding

Kokkinos [26] introduced a pioneering convolutional neural network, UberNet, that jointly handles multiple, low- mid- and high-level scene understanding tasks; boundary detection, surface normal estimation, saliency estimation, semantic segmentation, human part segmentation, semantic boundary detection, region proposal generation and object detection. The main contribution of this work was the training of a deep architecture while relying on diverse training sets for each task or group of tasks, and a limited memory budget. In [19], Liu et al. proposed the Multi-Task Attention Network (MTAN). This architecture consisted of a shared network, containing a global feature pool, and task-specific attention mechanisms at different feature levels. The task-specific attention modules determined which features are useful for which task by querying the global feature pool. However, the method is constrained to a few visual scene understanding tasks, namely, semantic segmentation and depth estimation.

In [96], the authors addressed depth estimation, surface normal prediction and semantic segmentation jointly. The method is dubbed Pattern-Affinitive Propagation network (PAPNet), where initially a shared feature extractor acted as the encoder. Three task-specific decoders followed to yield the initial predictions. To enhance multi-task interactions, the authors proposed to learn an affinity matrix to represent the pairwise multi-task relationships. A diffusion layer was introduced to refine the features using the affinity matrix as input. Thereafter, the refined per-task features were upsampled via three reconstruction networks to produce the task-specific outputs. Similarly to the previously discuss works, the authors addressed image segmentation, depth and surface normal estimation. In a similar approach [22], the authors investigated on semantic and human parts segmentation, and depth, saliency and surface normal predicton. This work argued that multi-task relationships are not scale agnostic. They demonstrated that tasks with high affinity at some scale do not necessarily retain their affinity at some other scale and vice versa. Using a standard feature extractor [1], the proposed network yielded the initial predictions at each image scale (resolution). Next, the authors employed spatial attention to distill knowledge from other tasks and refine the features of the initial predictions, across scales. Finally, the task-specific distilled features from all scales are aggregated to yield the final task outputs. Likewise, they addressed issues such as the limited effective receptive field of convolutional architectures at higher scales, by propagating information from the lower scales.

Recently, multi-task methods leveraged the potential of Neural Architecture Search to learn the optimal allocation of resources across tasks. Sun et al. [30] proposed Adashare, to address multi-task resource allocation by learning a select-or-skip policy for each task and each residual block of the network architecture. Following the common NAS paradigm, Adashare assumes independent blocks that can be skipped by some task should they not substantially contribute to its loss minimization. However, the method relies on an intensive bi-level optimization

scheme and a form of curriculum learning to learn the optimal architecture topology and its weights. The experimental results, again, revolved around similar, scene understanding tasks, i.e. semantic segmentation, surface normal, depth, keypoint and edge prediction. For the same tasks, but more recently, Bruggemann et al. [21] proposed ATRC, a convolutional multi-task architecture that enables cross-talk by learning different types of attention mechanisms for different tasks. The authors assumed that different pairs of tasks require different types of attention granularity i.e. some tasks might require global attention while for others local attention might be more optimal. In contrast to Adashare [30], ATRC trained with single-level optimization, i.e. both the architectural topology and model weights are trained with a single backward pass.

2.2.3 Visual Scene Enhancement and Understanding

Notably, only a very few works consider both enhancement and understanding tasks in the same architecture. Drawing on the relationship between motion and blur, Jung [97] introduced a motion-aware feature learning framework for dynamic scene deblurring using multi-task learning. Their approach simultaneously estimated a deblurred image and a motion field, leveraging a shared encoder architecture to effectively distinguish between various types of blur and improve motion estimation. However, motion estimation acted only as auxiliary task and the experiments were focused on deblurring. In another approach, Guo et al. [98], proposed a multi-task convolutional network, dubbed D3-Net, to perform deblurring, dehazing, and object detection within a single network. D3-Net employed image reconstruction and detection feature modules, to enhance image quality while detecting obstacles simultaneously. The authors adopted the visual scene enhancement tasks to improve the image quality and achieved higher obstacle detection performance. This work addressed obstacle - such as ships and bridges - detection for intelligent navigation in water scenes. Very recently, Nazir et al. [99] proposed to learn deblurring and depth estimation jointly with an adversarial multi-task network but did not consider video inputs.

2.2.4 Optimization

Despite their efficiency and – under conditions – improved performance, training multi-task networks involves the minimization of objectives exhibiting different magnitudes, angles and training speeds resulting in the dominance of some tasks over some others. A relevant optimization routine is illustrated in Fig. 2.1. However, when tasks are heterogeneous, with various objective functions, different tasks usually exhibit different gradient angles and magnitudes, resulting in suboptimal weight updates and uneven training rates across tasks.

Multiple multi-task optimization works (MTOs) have been proposed to alleviate the issue. Learnable balancing methods [32, 33, 34] involve learning parameters

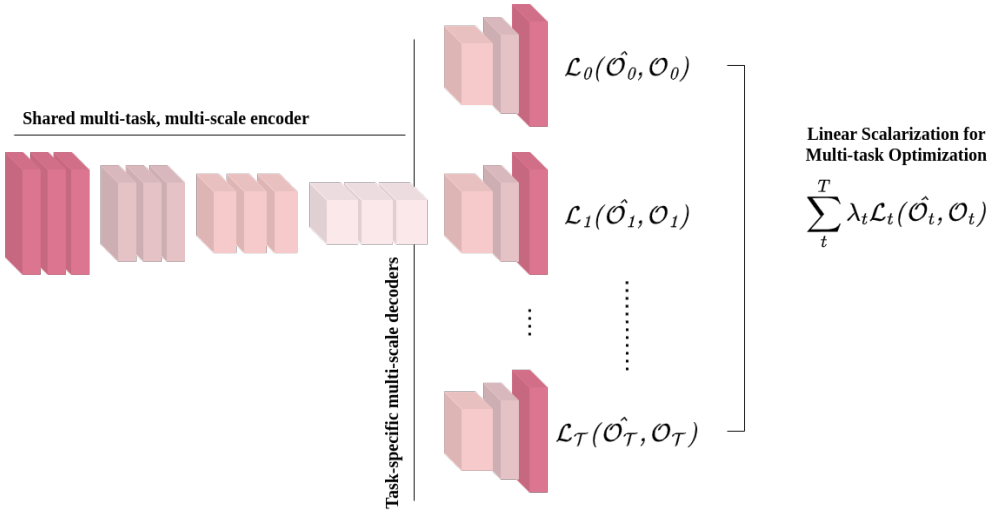


Figure 2.1: Linear scalarization, a typical multi-task optimization scheme in most relevant works, scales each per-task loss component with a weight.

to balance the gradients or the loss weights while adaptive balancing methods [19, 35] employ heuristics. Most MTOs require access to the per-task gradients to synthesize an enhanced weight update, but one major drawback is that memory and, thus, required time grows linearly with the number of tasks. This issue is shown in [33] and [37]. To further emphasize the MTOs inefficiency, [38] and [36] show that these methods can train up to 35 times slower than scalarization methods, all while failing to improve multi-task pareto curve trade-off. Simple baselines [39, 36], claim comparable performance without computational overhead. Recent research [38, 39, 36] casts doubt on the effectiveness of MTOs. [36] defends unitary scalarization with sufficient regularization. [39] propose random weight sampling to escape bad local optima, while [38] show that MTO methods fail to outperform traditional approaches despite added complexity and computational demands.

2.2.5 Datasets

Several general-purpose datasets have been developed to facilitate research in various areas, including object detection, classification, and semantic segmentation. The Common Objects in Context (COCO) dataset [100], for instance, contains a sizable collection of diverse images, with over 300,000 images spanning 80 object categories. This dataset is useful for tasks such as object detection, semantic segmentation, keypoint detection, and image captioning. However, not all images in COCO are annotated with both segmentation and keypoints, resulting in an asymmetrically annotated dataset. Another dataset that supports research in object detection, segmentation, and classification is PASCAL-VOC (Visual Object Classes) [101]. PASCAL-Context [24], a split of the larger, PASCAL-VOC dataset, provides a diverse collection of images in different scenes, with a total of over 10000 images labeled for semantic segmentation, human parts segmentation and boundary detection. Due to the research interest in multi-task learning, the dataset has been

augmented with saliency maps and surface normal labels [102], which are utilized in recent multi-task works [21]. NYUv2 [25] is another widely used dataset that provides labels for semantic segmentation, depth prediction, surface normal estimation, and boundary detection. It consists of over 1400 images of indoor scenes, each labeled with 13 classes of different objects. The dataset includes RGB images and corresponding depth maps recorded from the Microsoft Kinect. NYUv2 offers large variability in object sizes, occlusions, and lighting conditions, making it a valuable resource for the researche community.

More datasets have been proposed for multi-task learning, originating from the automotive domain and looking towards autonomous driving. The Cityscapes dataset [23] is a popular benchmark for depth estimation and semantic understanding of urban street scenes. It includes high-quality pixel-level annotations from high-resolution street-view images, recorded in 50 cities across different countries. It contains 5000 images, including people and vehicles, segmented across 30 different categories. The images are captured from different perspectives, with varying scenes, background and weather conditions. Later work [103] distilled the dataset with additional labels to detect whether objects are static or moving. Simiarly, The KITTI dataset [104] is a popular dataset used in the field of computer vision and machine learning, intended for autonomous driving and named after Karlsruhe, Germany, the city where it was collected. The dataset has contributed significantly to the advancement of the field since it comes with annotations for several tasks [105, 106] including but not limited to semantic segmentation, depth estimation, optical flow. However, these annotations were done separately for each task and the input is not always common across the tasks. In another such work, the Berkeley DeepDrive Industry Consortium introduced the BDD100K dataset [107], which represents a large-scale diverse driving video dataset with rich annotations. It consists of 100000 videos, with diverse weather and time-of-day conditions, and GPS/IMU information. This diversity is essential for testing the robustness of perception algorithms under various real-world scenarios. The dataset includes annotations for image tagging, object bounding boxes, drivable areas, lane markings, and full-frame instance segmentation. It is suitable for studying various aspects of autonomous driving, such as object detection, lane detection, and drivable area prediction.

2.2.6 Literature Gaps

While multi-task learning has attracted significant research interest, there exists a notable bias towards visual scene understanding compared to its enhancement counterpart. This bias can be attributed to various factors, such as the perceived financial viability of automating tasks related to scene understanding. Unlike enhancement, which is often seen as a facilitator for human decision-making, such as medical image diagnosis, or as a preliminary step to optimize performance for subsequent tasks, scene understanding promises direct automation and operational efficiency gains. However, similarly to the versatility of the human brain in pro-

cessing various types of visual information, the next generation of large-scale vision models should be capable of addressing a comprehensive set of vision tasks, including both understanding and enhancement ones. Such models will be capable of offering more holistic solutions, and address diverse application scenarios.

Although numerous successful single-task approaches have been developed for visual scene enhancement tasks such as dynamic video deblurring and denoising, there remains a notable gap in addressing both tasks concurrently. This thesis attempts to address this gap, by considering the simultaneous mitigation of blurring and noise in various settings, from lower to heavier levels. To tackle this challenge, we propose the first multi-task learning framework for video deblurring and denoising, that exploits the information present in video footage to enhance performance. Thereafter, we merge video deblurring with semantic segmentation to address enhancement and understanding tasks jointly. Extending our work further, later in this thesis, we merge deblurring and denoising with color mapping, homography estimation and semantic segmentation. Effectively, our work accommodates a multitude of diverse tasks in competitive and runtime-efficient architectures.

Although numerous datasets exist for multi-task learning, they mainly focus on tasks related to understanding visual scenes, with similar tasks across datasets, differing only in size and scene content. During this doctoral dissertation, we investigated existing datasets and label generation processes. As a result, we constructed two new datasets, one available to the public now and one under review, soon-to-be released. These datasets are unique since they cover a wide range of vision tasks, spanning from video denoising, deblurring and color mapping to homography estimation and semantic segmentation. Situated in the field of medical imaging, our datasets are accompanied by very competitive and efficient multi-task methods, rendering them invaluable testbeds for evaluating algorithmic performance of other works.

While exploiting multi-task interactions proves advantageous for multi-task learning, extending this approach to the spatiotemporal domain offers even richer information sharing among tasks. Despite its potential, only a handful of studies have explored the use of video footage for multi-task learning. In the one of the few works published while authoring this doctoral thesis, Chennupati et al. [108] introduced a siamese encoder that aggregates features from various layers and assigns them to different tasks using separate decoders. However, this method of frame fusion has limitations. Specifically, it disregards feature alignment and fuses frames directly, i.e. with simple concatenation of their features. Earlier works, especially on the visual scene enhancement domain, used optical flow [71, 109, 110]. However, more recent research [50, 62, 111] has highlighted the drawbacks of optical flow methods when not jointly trained with restoration tasks. In this thesis, we propose multi-task learning architectures that employ learnable kernels [112, 50] for feature alignment, deformable convolutions [58, 113, 62], and homographies [114, 115, 116]. All our methods incorporate feature alignment modules trained end-to-end with primary tasks. Additionally, within the scope of this thesis, we conducted another

study focusing solely on video denoising, i.e. in a single-task manner. There, we utilize multi-scale features to implicitly address motion-related challenges.

Chapter 3

Video Multi-task Learning of Low- and Mid-level Tasks for Visual Scene Enhancement

3.1 Introduction

In this Chapter, we address the first component of our first hypothesis. We jointly address two visual scene enhancement tasks in RGB video scenery, Videos aim at faithfully reflecting the motion in dynamic scenes but concurrent motion blur and noise can severely obscure scene perception. To reduce artifacts, one could calibrate the camera sensor. Proper calibration requires adjustment of the exposure time. While a longer time of exposure increases the number of photons and thus allows the sensor to capture scenes with less noise, it increases the risk of motion blur when the camera shakes and objects move. However, a small exposure time causes noise. Therefore, the two degradation factors are interrelated. As follows naturally, this relationship can be leveraged in concurrent video denoising and deblurring. Despite the well-studied noise-blur trade-off introduced on the optics level, and the affinities between the tasks, the two visual scene enhancement tasks have not been addressed jointly. In this Chapter, we investigate the first component of our first hypothesis, i.e. the optimal approach to merge two foundational visual enhancement tasks such as video denoising and deblurring,

The problem at hand raises questions. *Should different models be tailored to individually address denoising and deblurring tasks or is a single, multi-task model a more efficient approach? How robust are current deep video restoration methods with increasing noise levels? Is there an architectural topology capable of leveraging i) motion and ii) the inter-task affinities to efficiently address joint video denoising and deblurring?*

Improving visual outputs finds applications in visualization environments where the user can assess the scene more accurately and react. Moreover, enhanced video

processing facilitates downstream computer vision tasks and improves performance in general video understanding. Although algorithms should address hardware limitations and account for adversarial physical phenomena by enhancing the video output, satisfying the objectives of a real-world application is a demanding task in practice. Numerous methods have been proposed to address the deblurring task, ranging from spatially invariant [117, 118, 119, 120, 121, 122] to spatially variant blur [123, 54, 110, 50, 124, 48]. Meanwhile, many approaches have been proposed for denoising with remarkable results [46, 71, 125, 126]. However, deeply learnt, dynamic scene video denoising and deblurring have been addressed only as independent tasks. The problem of spatially variant motion blur in the presence of noise, has not yet been addressed in the deep learning literature. In this Chapter, we use real-world blurry dataset of [48] and the realistic Poisson-Gaussian noise model [127, 128]. Our application focuses on real-world outdoor scenes but further tests our findings on a novel, dental dataset.

This Chapter discusses the first, deeply-learnt network that leverages the feature-sharing potential of multi-task learning (MTL) to increase model efficiency and jointly address dynamic video denoising and deblurring. Intially, we propose *R2-D4*, a novel, MTL-inspired, cascaded convolutional architecture utilizing two decoders to denoise and deblur input frames in stages. *R2-D4* uses an alignment module that leverages deformable convolutions at the feature level. Thereafter, we introduce multiscale residual dense modules to learn coarse-to-fine, dense representations, enhanced by MECA, a novel extension of the efficient channel attention module [129] to further modulate deformable convolutions and increase restoration performance while retaining the number of FLOPs. We extensively benchmark existing deblurring approaches under different levels of noise on a real, publicly available dataset and show that state-of-the-art deblurring networks bear noise-removing capacity, yet *R2-D4* performs consistently better.

3.2 Method

Let us assume a video camera, which streams frames B_t at each time step t. For each frame B_t, let x correspond to pixel location. Given pixelwise blur kernels k_t of size K, the degraded image is generated as:

$$B_t = \sigma_n(R_t * k_t)(x) + \eta_n \tag{3.1}$$

Here, $*$ denotes the convolution of R_t with the blur kernel k_t at x, η_n represents additive noise, and σ_n stands for signal-dependent noise. However, we can view the convolution of R_t with k_t as the inner product of R_t with $\tilde{k}_t$ translated by x. This is equivalent to:

$$B_t = \sigma_n\langle R_t, T_x\tilde{k}_t\rangle + \eta_n, \tag{3.2}$$

where $\tilde{k}_t$ denotes the involution with $\tilde{k}_t = k_t(-x)$ and T_x refers to the aforementioned translation component. However, the kernel weights of convolutional neural networks are learnable and thus the cross-correlation is typically utilized for notational convenience. Then, a simple inner product can be used, without involution:

$$B_t = \sigma_n \langle R_t, T_x k_t \rangle + \eta_n \tag{3.3}$$

If we limit the operations locally, the translation component T_x is ommited. Now, $(R_{x,t})^-$ refers to a window of size K centered around pixel x in the image R_t and $B_{x,t}$ is the resultant window on the blurry image. Likewise, the convolution at each input pixel location is defined as:

$$B_{x,t} = \sigma_n \langle (R_{x,t})^- k_{x,t} \rangle + \eta_n, \tag{3.4}$$

In this Chapter, we are interested in obtaining the noise- and blur-free frames, sequentially. Following the inverse of Equation 3.4, we approach the problem by removing the noise first to obtain $\hat{R}'_t$, and subsequently the blur, to obtain the predicted clean frame $\hat{R}_t$.

3.2.1 Architecture

Overview

The proposed architecture is illustrated in Figure 3.1. Given N consecutive corrupted frames $B_{[t-N:t]}$ and $N-1$ previously restored frames $\hat{R}_{[t-N:t-1]}$, our method obtains $\hat{R}_t$ via a cascaded, two-stage restoration. The proposed *R2-D4* network consists of a shared, dense, deformable (D2) feature alignment module, followed by a convolutional feature fusion and two decoders performing denoising and deblurring sequentially (D2) to restore the frames via a two-stage (R2) cascaded process, as illustrated in Figure 3.1. The shared D2 module processes the frames at time steps. For each incoming frame at time step t, to predict $\hat{R}_t$, it takes as inputs current and past information. From the current timestep it requires the corrupted frame B_t, whereas for the past N-1 time steps, it accesses, in pairs, both the past corrupted $B_{[t-N:t-1]}$ and restored frames $\hat{R}_{[t-N:t-1]}$. At each time step t it extracts features $f_{[t-N:t]}$. Subsequently, the asymmetric offsets are estimated to align the neighboring frame features $f_{[t-N:t-1]}$ with the reference frame features f_t. Thereafter, two aligned sets of features are fused before the two decoders leverage the shared features to denoise and deblur the current frame sequentially in a cascaded manner.

Our D2 alignment module, employs modulated deformable convolutions[113] for feature-level frame alignment. Notably, it does not estimate optical flow, which is particularly challenging under strong noise and can introduce computational inefficiencies and motion artifacts. Feature alignment is further improved using our

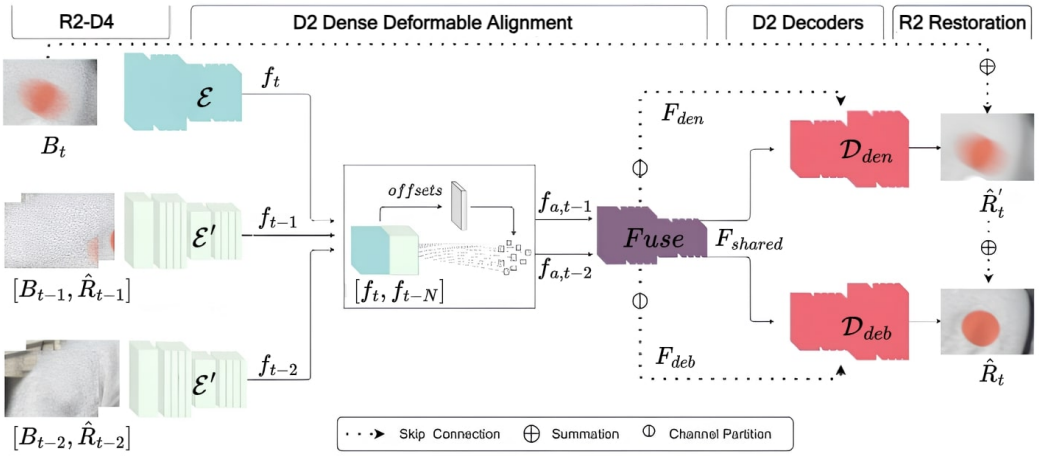


Figure 3.1: The proposed *R2-D4* architecture restores the reference frame (R2) via cascaded denoising and deblurring (D2) after aligning its features with the neighboring ones via the dense deformable (D2) alignment module.

multiscale residual dense modules (MS-RDMs). These modules leverage dilated convolutions to capture longer-range context, enhancing feature alignment by aggregating features with larger receptive fields. This helps address the deformable offset estimation issue[62, 130]. Furthermore, MS-RDMs are enhanced with our modulated efficient channel attention blocks, as explained in Sec. 3.9.

The *R2-D4* restoration process, consisting of denoising and deblurring decoders, operates sequentially under an efficient multi-task framework. Accurate feature alignment benefits both denoising and deblurring tasks. Additionally, channel-wise expansion of features upon fusion increases the model's capacity at lower resolutions, effectively accommodating both tasks. This two-stage cascaded restoration process has shown improved performance across various restoration tasks, leading to its integration into *R2-D4* through the proposed feature-sharing scheme. As depicted in Figure 3.1, we incorporate additional residual connections from B_t to the first-stage output $\hat{R}t'$ and from the latter to the second-stage $\hat{R}t$ to facilitate training.

Proposed Block: Modulated Efficient Channel Attention

First, we propose a novel channel attention block to integrate within *R2-D4*. Self-supervised channel attention blocks have become ubiquitous since they highlight informative and suppress non-relevant features. Wang et al. [129] proposed an efficient 1D convolution (ECA) on globally averaged input channels to determine the attentive weights, as illustrated in the top half of Figure 3.2. Formally, using the "composition" $\circ$ notation, ECA is denoted as follows:

$$ECA_k = \sigma \circ \mathcal{C}_{c,1\times k} \circ GAP \tag{3.5}$$

where σ is the sigmoid function and $\mathcal{C}_{c,1\times k}$ is a 1D convolutional operation with c output channels a kernel of size k, no bias, and padding to retain the dimensionality. GAP denotes the global average pooling operation. Formally, let us assume some feature cube f_t^s, for some time step t and feature resolution s, the channel-wise attention weights are derived as follows:

$$\tilde{w}_{att} = ECA_k(f_t^s) \tag{3.6}$$

Then, $\tilde{w}_{att}$ is multiplied by the input features f_t^s to obtain the attended F_t^s as follows:

$$F_t^s = \tilde{w}_{att} \times f_t^s \tag{3.7}$$

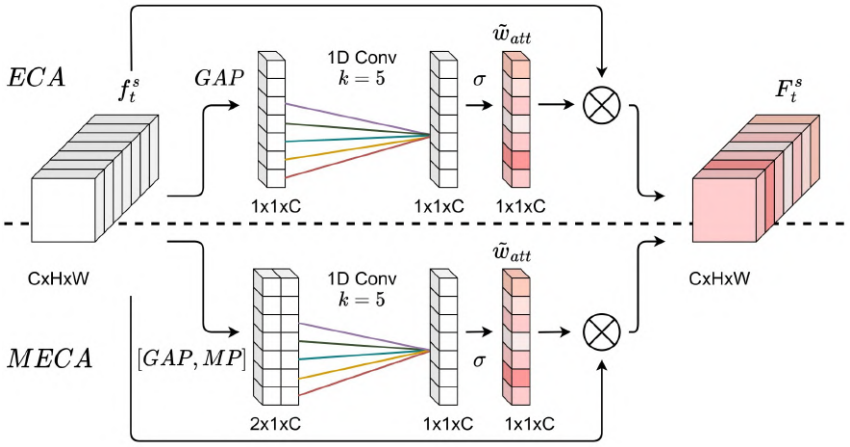


Figure 3.2: Proposed modulated efficient channel attention.

Despite the success of channel attention modules, they are often difficult to optimize and converge to uniform distributions of the channel weights. To alleviate such issues and facilitate the gradient flow during the backward pass, we propose to complement globally averaged features with max-pooled features as in CBAM [131], under the efficient 1D convolution configuration of [129]. The modulated efficient channel attention module, termed MECA, is illustrated in the bottom half of Figure 3.2. In contrast to ECA, we perform both global average and max pooling (MP) on the features f_t^s channel-wise to obtain $\tilde{w}_{att}$, and we denote the concatenation of GAP and MP channels as MGAP. By adopting the notation in Equation 3.5, MECA is defined as:

$$MECA_k = \sigma \circ \mathcal{C}_{c,2\times k} \circ MGAP \tag{3.8}$$

The attended weights are derived similarly to Equation 3.6,

$$\tilde{w}_{att} = MECA_k(f_t^s) \tag{3.9}$$

and multiplied by the input features to obtain the attended features:

$$F_t^s = \tilde{w}_{att} \times f_t^s \tag{3.10}$$

Notably, MECA retains the efficiency of 1D convolution in capturing the local cross-channel interactions but learns an essentially more effective projection, utilizing two channels of informative cues instead of solely the globally averaged ones. MECA is an easy-to-plug module that can be integrated into all standard architectures for any vision task.

Proposed Block: Multiscale Residual Dense Module

Subsequently, we propose multiscale residual dense modules to learn better feature representations. Residual blocks (RBs) [55] have been a popular choice [50, 126, 132, 133] in image and video restoration. More recently, residual dense blocks (RDBs) [134] exploited dense connections between layers to extract richer hierarchical features while instantiating a contiguous memory (CM) mechanism to further enhance the learned representations.

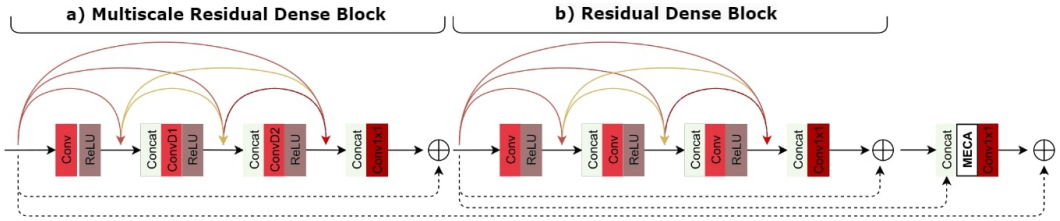


Figure 3.3: The proposed Multiscale Residual Dense Module learns enhanced hierarchical representations via its coarse-to-fine design. The MS-RDB (a) block mines coarser features with increasing dilation rates whereas the second RDB (b) block learns finer details.

RDBs typically consist of l convolutional kernels and a "growth factor" hyperparameter g. As shown in Figure 3.3b, each layer receives the feature maps from the previous stage, convolves them with a 3×3 kernel that yields g additional channels and concatenates them with the previous ones before passing them to the next layer. Each block is then followed by a 1×1 convolution to aggregate the signal and stabilize the training before the residual summation. Formally, a single RDB with 3 layers can be denoted as follows:

$$\begin{aligned} \mathcal{RDB}_{c,g} = & \mathcal{C}_{c,1\times1} \circ \mathcal{CAT}_{c+3g} \circ \mathcal{C}_{g,3\times3} \circ \\ & \mathcal{CAT}_{c+2g} \circ \mathcal{C}_{g,3\times3} \circ \mathcal{CAT}_{c+g} \circ \mathcal{C}_{g,3\times3} \end{aligned} \tag{3.11}$$

where $\mathcal{C}_{c,k\times k}$ and $\mathcal{CAT}_c$ are the $k \times k$ convolution operation and the concatenation function respectively. The subscript c denotes the number of output channels after each convolution and concatenation. Stacking b such residual dense blocks gives rise to RDB cells [48], where the output of each block is sequentially processed by the next block. For clarity, we term them residual dense modules (RDMs). In RDMs, all subsequent RDB outputs are concatenated and fed into another 1×1 convolution before the residual summation at the module level.

In this Chapter, we propose multiscale residual dense modules (MS-RDMs) to efficiently increase the effective receptive field by spatially augmenting the hierarchical features in a coarse-to-fine manner. As illustrated in Figure 3.3, MS-RDMs are designed via an MS-RDB that captures a hierarchically coarser context via kernel dilation followed by a simple non-dilated RDB to complement hierarchical features with fine details. Regarding the MS-RDB, layers are progressively enhanced with larger dilation rates to hierarchically capture a longer-range context. As depicted in Figure 3.3a, the MS-RDB block is defined as:

$$\begin{aligned}\mathcal{RDB}_{\mathcal{MS}c,g} =& \mathcal{C}_{c,1\times1,0} \circ \mathcal{CAT}_{c+3g} \circ \mathcal{C}_{g,3\times3,2} \circ \\ & \mathcal{CAT}_{c+2g} \circ \mathcal{C}_{g,3\times3,1} \circ \mathcal{CAT}_{c+g} \circ \mathcal{C}_{g,3\times3,0}\end{aligned} \tag{3.12}$$

where $\mathcal{C}_{c,k\times k,d}$ denotes, again, the convolution, but dilated with a rate of d. Upon concatenation of the coarse and fine block features and before the 1×1 convolutional aggregation, we perform channel-wise attention via the proposed $MECA_k$. Similarly, the resultant RDM_{MS} is defined as:

$$\begin{aligned}\mathcal{RDM}_{\mathcal{MS}c,g} =& \mathcal{C}_{c,1\times1} \circ MECA_7 \circ \\ & \mathcal{CAT}_{2c} \circ \mathcal{RDB}_{c,g} \circ \mathcal{RDB}_{\mathcal{MS}c,g}\end{aligned} \tag{3.13}$$

The proposed MS-RDM reformulation enlarges the effective receptive field, which in turn renders the CM mechanism spatially more aware. The coarse-to-fine hierarchical features mine spatially aware representations and serve as a preprocessing step for deformable offset estimation.

Cascaded Restoration: Dense Deformable Alignment

The proposed architecture is illustrated in Figure 3.1. At each new time step t, R2-D4 takes as input the current corrupted frame B_t and $N-1$ previously corrupted $B_{[t-N:t-1]}$ and restored frames $\hat{R}_{[t-N:t-1]}$. Leveraging previously restored frames encourages temporal coherence by reducing flickering and has been shown to yield improved performance [50]. At each time step, the respective features are computed using the following block:

$$\mathcal{E} = \mathcal{RDM}_{\mathcal{MS}32,32} \circ \mathcal{C}_{32,3\times3,2} \circ \mathcal{RDM}_{\mathcal{MS}16,16} \circ \mathcal{C}_{16,3\times3,1} \tag{3.14}$$

where $\mathcal{C}_{c,k,s}$ denotes a $k \times k$ convolution with a stride of s, and c output channels and $\mathcal{RDM}_{\mathcal{MS}c,g}$ is the multiscale residual dense block with a growth factor g and, again, c output channels. As illustrated in Figure 3.1, *R2-D4* contains two identical blocks: $\mathcal{E}$ for the current B_t and $\mathcal{E}'$ for each past $\{B_{t-N}, \hat{R}_{t-N}\}$. Likewise,

$$f_t = \mathcal{E}(B_t), \qquad f_{t-N} = \mathcal{E}'(B_{t-N}, \hat{R}_{t-N}) \tag{3.15}$$

where N is experimentally set to 2 past frames. Weight sharing for past frame features increases the training efficiency and accelerates inference by reusing f_{t-2} at each time step.

The current and previous frames are then aligned using deformable convolutional layers. A deformable module enables the modeling of geometric transformations through asymmetric kernels so that output features can capture object-specific contexts that assist blur kernel estimation. The leveraging of deformable convolutions under the proposed scheme has three advantages. First, it discards the necessity for erroneous and computationally expensive optical flow estimations. Second, it performs alignment on the deeper feature levels instead of the image level. This has been shown to improve performance [50, 62] because the layers prior to the deformable modules encode features that are tailored to the alignment. Third, estimating deformation offsets on the coarse-to-fine features extracted from MS-RDMs assists in modulated offset estimation and improves performance.

Each modulated deformable layer consists of two convolutions. The first layer learns the offset displacements and the modulating scalars that determine the amplitude of the output features. The second layer employs the modulated offsets and learns the filter weights, as in ordinary convolution. The deformable convolution is denoted as

$$\mathcal{DC} = \mathcal{C}^{D}_{128,3\times3,1} \circ \mathcal{C}_{27,3\times3,1} \tag{3.16}$$

where $\mathcal{C}_{3k^2,k\times k,s}$ is the $k \times k$ convolutional kernel with a stride of s, estimating the $2k^2$ offsets and respective k^2 modulation scalars from the concatenated c frame features and $\mathcal{C}^{D}_{c,k\times k,s}$ denotes the actual deformable convolution with c output channels. Correspondingly, the aligned features are defined as follows:

$$f_{a,t-N} = \mathcal{DC}(f_t, f_{t-N}) \tag{3.17}$$

The fusion of the aligned features is then performed via:

$$\begin{aligned} Fuse = \mathcal{RDB}_{32} \circ \mathcal{RDB}_{32} \circ \mathcal{C}_{128,3\times3,2} \circ \\ \mathcal{RDB}_{32} \circ \mathcal{RDB}_{32} \circ \mathcal{C}_{128,3\times3,1} \end{aligned} \tag{3.18}$$

Note that simple RDBs without dilation rates are employed for fusion because a spatially wider context does not strengthen the feature representations at smaller scales. Because the number of past frames is $N = 2$, the output and shared features are defined as

$$F_{shared} = Fuse(f_{a,t-1}, f_{a,t-2}) \tag{3.19}$$

Cascaded Restoration: Decoders

The decoders share identical architectures. They are optimized to upsample the shared features and yield denoised and deblurred outputs (D2) sequentially. As shown in [135], transposed convolutions often generate checkerboard artifacts. To

overcome these problems during feature upsampling, many studies have resorted to bilinear upsampling followed by convolution [136, 60]. Although we confirm that bilinear upsampling eliminates artifacts, it leads to a loss of spatial information. Therefore, we resort to convolutional channel-wise expansion followed by pixel shuffling [137] to reduce gridding artifacts and preserve spatial details. Denoting the upsampling layers as PS, each decoder can be expressed as follows:

$$\begin{aligned}\mathcal{D} = \mathcal{C}_{3,3\times3,1} \circ \mathcal{RDB}_{32,16} \circ \mathcal{PS}_{32} \circ \mathcal{C}_{128,3\times3,1} \circ \\ \mathcal{RDB}_{64,32} \circ \mathcal{PS}_{64} \circ \mathcal{C}_{256,3\times3,1}\end{aligned} \tag{3.20}$$

Assuming two such instantiations for denoising and deblurring as $\mathcal{D}_{den}$ and $\mathcal{D}_{deb}$, the intermediate denoised and restored output frames are defined as:

$$\hat{R}_t^{'} = \mathcal{D}_{den}(F_{shared}) \tag{3.21}$$

$$\hat{R}_t = \mathcal{D}_{deb}(F_{shared}) + \mathcal{D}_{den}(F_{shared}) \tag{3.22}$$

We further utilize skip connections from the encoder to the decoders to preserve spatial information and facilitate training, as is common in UNet-based [73] methods. Instead of concatenating the encoding channels with both decoders, we restructure the gradient flow by dissecting the former, say $F \in \mathbb{R}^{H\times W\times C}$, in two groups $F_{den}, F_{deb} \in \mathbb{R}^{H\times W\times C/2}$, each specialized for the decoder's task, as illustrated in Figure 3.1. Likewise, F_{den} and F_{deb} receive task-specific gradients in addition to the shared gradients. As a result, F_{den} focuses on the global noise distribution, whereas F_{deb} is specialized in recovering the blur-free frame.

3.2.2 Loss Function

The *R2-D4* parameters are derived by optimizing the following objective, where $\mathcal{L}$ is a weighted sum of ℓ_2 squared norms, i.e.

$$\mathcal{L} = \mathcal{L}_{blur} + \lambda_1 \mathcal{L}_{noise} + \lambda_2 \mathcal{L}_{perceptual}, \tag{3.23}$$

where

$$\mathcal{L}_{blur} = \frac{1}{CHW} ||R_t - \hat{R}_t||^2, \tag{3.24}$$

$$\mathcal{L}_{noise} = \frac{1}{CHW} ||R_t^{'} - \hat{R}_t^{'}||^2, \tag{3.25}$$

and

$$\mathcal{L}_{perceptual} = \frac{1}{C_\phi H_\phi W_\phi} ||\phi_{VGG}(R_t) - \phi_{VGG}(\hat{R}_t)||^2. \tag{3.26}$$

$\hat{R}_t^{'}$ and $\hat{R}_t$ are the "only denoised" and "fully restored" cascaded model outputs, while $R_t^{'}$ and R_t are the respected ground truth labels. The definition of $\mathcal{L}_{perceptual}$ is adopted from [138], where ϕ_{VGG} denotes the VGG-19 features [3] extracted from the 3th layer and C_ϕ, H_ϕ, W_ϕ denote the corresponding feature dimensions. The scalar

values C, H, and W refer to the image channel, height, and width, respectively, and the weights are experimentally set to $\lambda_1 = 0.6$ and $\lambda_2 = 0.01$.

3.3 Experiments

3.3.1 Dataset

All experiments use the "3ms24ms" version of BSD that has the strongest level of blur. The evaluation protocol contains 60 training (30K pairs), 20 validation (10K pairs) and 20 test (15K pairs) sequences with a resolution of 640×480. The BSD dataset therefore already contains the pairs of blurry and clean frames $(R_t^{'}, R_t)$. We need to further generate noise on top of $R_t^{'}$ and obtain the both noisy and blurry frames B_t. Following Equation 3.1, we generate Poisson-Gaussian noise directly on the blurry frames of the BSD dataset. Specifically, for each input blurry pixel z, we generate noise as $x = az(y/a) + w$ where x denotes the both blurry and noisy pixel. Here, a is a scaling parameter which controls the Poisson noise while $w \sim N(0, \sigma^2)$ dictates the strength of Gaussian noise. The aforementioned, pixel-wise operation is applied on whole frames to generate the corrupted frames B_t. Consequently, we obtain the triplets $(B_t, R_t^{'}, R_t)$ required to train the cascaded deep network for sequential restoration.

To simulate different noise levels we generate noise under three different sets of parameters $\{\alpha, \sigma\}$ equal to {0.5, 0.9}, {1.9, 1.7} and {7.1, 3.3} for low, moderate, and severe noise, respectively. The generated shot noise distribution is typical for bright, dark, and low-light images for α equal to 0.5, 1.9, and 7.1, respectively.

Last, we further assess the transferability of our model and showcase qualitative results of the proposed model on an in-house dataset of natural teeth.

3.3.2 Setup

In this section, we outline the experiments designed to achieve the following objectives: (i) demonstrate the effectiveness of a deep learning model cascade, (ii) compare the performance of *R2-D4* with state-of-the-art video deblurring methods across various noise levels, (iii) evaluate the impact of the proposed architectural enhancements in *R2-D4*, (iv) assess the influence of the multi-task learning configuration, and (v) analyze the effects of model pruning on performance. To address these goals, we conducted the following experiments:

- Sequential Methods Comparison: Initially, we evaluated the performance of a basic system in which two methods operated sequentially. We trained FastDVDNet [46] for denoising, followed by *STFAN* [50] for deblurring.
- Comparison with State-of-the-Art Models: We compared *R2-D4* against state-of-the-art models, including *STFAN* [50], *ESTRNN* [48] with 15 blocks using

only past frames, and *CDVD-TSP* [63].

- MTL Ablation Study: To assess the effectiveness of our multi-task learning setup, we compared it with R2-D3, which uses a single decoder.
- Impact of Proposed Blocks: We examined the influence of the proposed blocks on our feature alignment module by considering *R2-D4*$^-$, which involves: (i) replacing MECA with ECA, and (ii) substituting MS-RDB modules with simple RDB modules while retaining GFLOPs.
- Model Pruning: We explored reduced variants of *R2-D4* by reducing the number of channels in the decoders and the fusion module, resulting in "small" and "medium" *R2-D4* variants that are more computationally efficient.

Our experiments were conducted using PyTorch on an Nvidia Tesla V100 GPU, spanning 250 epochs. The weights were set to $\lambda_1 = 0.6$ and $\lambda_2 = 0.01$. The Adam optimizer [139] was employed, with a learning rate initialized at 1.5×10^{-4} and decayed to 10^{-6} using the cosine annealing strategy [140]. The networks were trained on sequences of 30 frames with a batch size of 1. For consistency, experiments involving state-of-the-art methods adhered to their official publicly available implementations, and followed the same data augmentation strategies [50] within a unified codebase. We evaluate performance with PSNR and SSIM [141], as the most widely adopted metrics for visual scene enhancement.

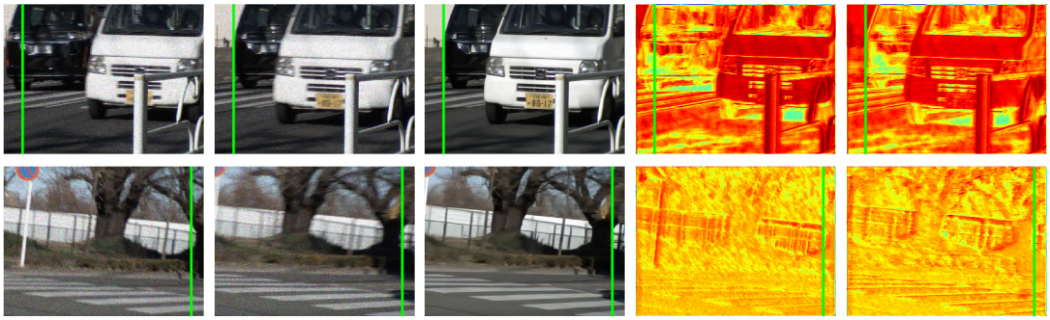
3.3.3 Results

The naive approach is not trained end-to-end and thus oversmooths the input frames achieving a PSNR of 28.40 and an SSIM of 0.850 for the severe noise setting. The results of the end-to-end methods are listed in Table 3.1. Interestingly, our experiments show that deblurring methods bear some noise-removal capacity, although *R2-D4* performs better than *STFAN* and *ESTRNN* in both PSNR and SSIM. Moreover, it performs higher in PSNR and on par in SSIM with the computationally expensive, cascaded version of *CDVD-TSP* (2), which performs two passes over the corrupted frames and uses five input frames. As shown in Table 3.1, the performance increased over the compared methods across all levels of noise. The second decoder and the proposed blocks clearly contribute to performance gains, increasing mean PSNR by 0.19 dB and 0.15 dB compared to R2-D3 and *R2-D4*$^-$, respectively. Last, Figure 3.5 shows that while the small *R2-D4* variant has 30% fewer GFLOPs in comparison to *ESTRNN*, it performs better than both *STFAN* and *ESTRNN*.

R2-D4 benefits from accurate feature alignment under strong noise and recovers fine-grained frame details (see Figure 3.6). One can observe that *STFAN* often fails to align features producing hallucinations, as seen in the gas tube (top row) and in the fence (middle row). For the same examples, the *ESTRNN* tends to oversmooth the output. *CDVD-TSP* performs better but tends to yield piecewise

Architecture	N	# P(M) ↓	GFLOPs ↓	Low ↑	Moderate ↑	Severe ↑
STFAN	2	5.4	188.9*	29.23	29.06	28.57
				0.875	0.868	0.858
ESTRNN	3	2.3	142.9	30.52	29.90	29.07
				0.905	0.892	0.872
CDVD (1)	5	16.2	-	30.40	29.92	29.06
				0.906	<u>0.894</u>	0.875
CDVD (2)	5	16.2	-	30.53	30.12	29.17
				0.911	**0.900**	**0.880**
R2-D3	3	4.4	216.9	30.82	30.19	29.17
				0.905	0.886	0.870
R2-D4^{-}	3	5.1	270.7	<u>30.93</u>	<u>30.22</u>	<u>29.18</u>
				0.907	0.890	0.870
R2-D4	3	5.1	270.7	**31.10**	**30.32**	**29.33**
				<u>0.910</u>	<u>0.894</u>	/ <u>0.876</u>

Table 3.1: Results of the proposed methods compared to state-of-the-art single-task solutions on the test. PSNR (top) and SSIM (bottom) results at three noise levels are illustrated. GFLOPs* for STFAN did not include their FAC layers. The bold and underlined results indicate the first and second rank, respectively.



(a) Input B_{t-2} (b) Input B_t (c) Output $\hat{R}_t$ (d) Offsets $t-2$ (e) Offsets t

Figure 3.4: Visualization of deformable offsets. *R2-D4* adapts the offsets for independently moving or uniform motion scenarios.

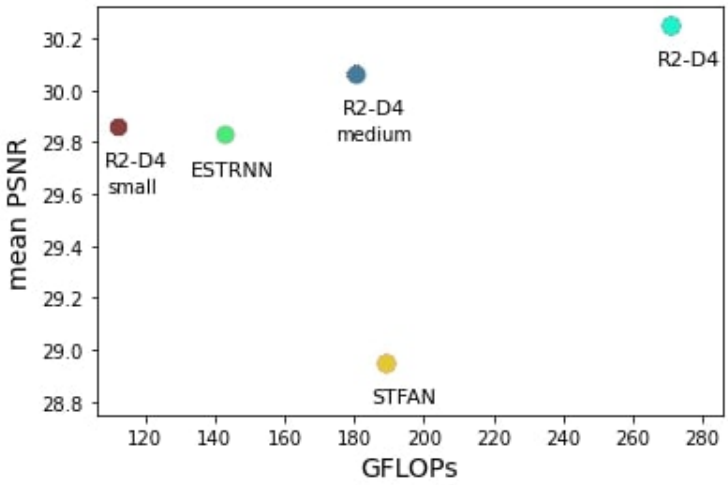


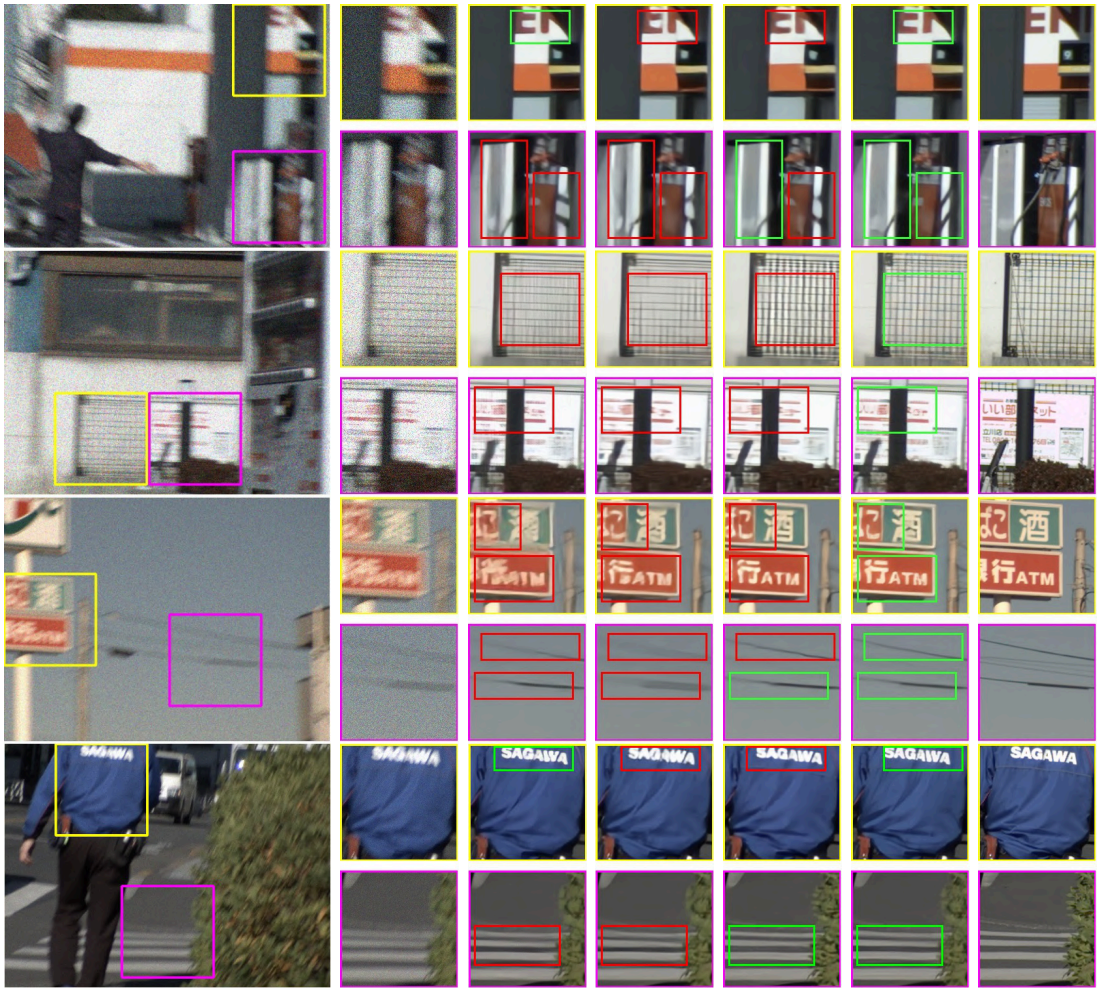
Figure 3.5: Mean PSNR versus GFLOPs for three *R2-D4* variants compared to *ESTRNN* and *STFAN*.

constant artifacts despite its larger complexity, which is visible in the fence example. *R2-D4* performs implicit feature alignment and dynamically adapts offsets over time, as illustrated in Figure 3.4. The top row illustrates the scenario of independently moving objects, whereas the bottom row depicts the uniform motion caused by camera movement. The offset variance is higher for the former; *R2-D4* mines the spatio-temporal boundaries and aggregates the object-specific context. The spatial responses for the second case show a smaller variance as the learned offsets exhibit similar directions. *R2-D4* dynamically adapts offsets in the case at hand.

Moreover, we employ a in-house dataset with paired, noisy and blurry video sequences, and showcase how our R2D4 model not only denoises and deblurs but also transforms the image colors so that they are vivid and visually more pleasant. As shown in Figure 3.7, our model is able to effectively not only remove the noise and change the color, but also significantly reduce the amount of blur present in the input frames.

3.4 Discussion and Conclusion

In this Chapter, we studied two visual enhancement tasks, dynamic scene video deblurring and denoising under strong noise. Although such acquisition settings arise frequently in practice, the problem is challenging and new in the deep learning literature. The main challenges lies in the fact that their outputs are less correlated. While the noise map is omnipresent on the whole image, deblurring in dynamic scenes with diverse depths is heavily space-variant. We demonstrated that state-of-the-art deblurring methods have some denoising capacity. Yet, the proposed *R2-D4* method outperformed them owing to a multi-task, cascaded yet efficient architecture, enhanced with multiscale residual dense modules. We showed that we can achieve state-of-the-art performance on this challenging combination of visual



(a) Input (b) Patch (c) STFAN (d) EST (e) CDVD (f) R2D4 (g) GT

Figure 3.6: Qualitative Results of R2D4 against compared methods. In zoomed areas, red and green rectangles highlight artifacts and more accurate reconstructions, respectively. The first, second, third and fourth rows were generated with severe, severe, moderate and low noise.

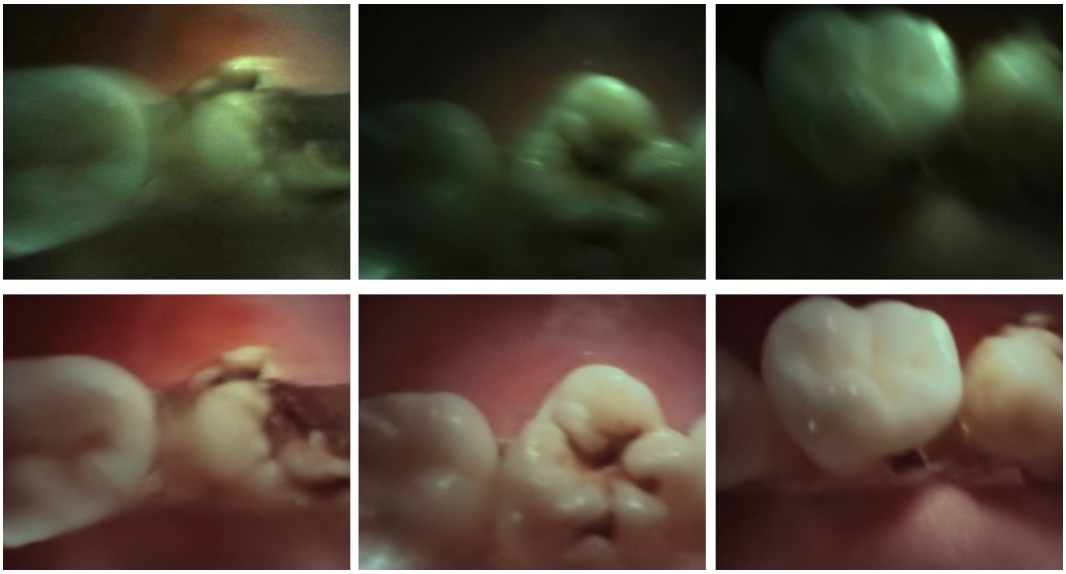


Figure 3.7: Qualitative results on an in-house dental dataset. Clearly, R2D4 generalizes across scenes and is able to restore video frames in multiple application environments.

scene enhancement tasks, even with very high levels of noise.

The performance improvement over compared methods originated from the architectural design as we show that even with lesser number of parameters and floating point operations, *R2-D4* still performed better than end-to-end, single task networks. To a large extend, we posit the success of this architecture to the deformable offsets which function as accurate and robust motion predictors. The strong supervision from the multi-task signal was in itself enough to learn pixel displacements, align frames and effectively borrow context from neighbouring ones to restore even heavily degraded frame regions. This Chapter answers the first component of our first hypothesis; the proposed architecture can tackle both tasks with higher performance and reduced compute compared to single-task networks. Last, we verified that our proposed model is robust to various channel pruning configurations and capable of generalizing across different video scenes. This Chapter includes in part work published in [142]

Chapter 4

Video Multi-task Learning of Mid-level Tasks for Dental Scene Enhancement and Understanding

4.1 Introduction

In this Chapter, we investigate the second component of our first hypothesis and study two diverse tasks, i.e. video motion deblurring and object segmentation. The tasks of video deblurring and segmentation are more closely interconnected. Motion blur is particularly noticeable, especially at the edges of moving scenes and objects. Furthermore, the process of deblurring a video frame can significantly enhance the accuracy of object segmentation. Since motion blur serves as a vital visual cue, detecting blur can offer valuable insights into the segmentation layers within dynamic, moving scenes [143]. Previous research has demonstrated that blur detection can provide partial information about the segmentation process in moving scenes [144],[145],[146],[147]. Additionally, image segments can play a pivotal role in guiding the estimation of blur kernels, which model the varying motion patterns. These segments indicate object boundaries that should remain sharp during the deblurring process. Accordingly, we investigate an architecture capable of accommodating both tasks jointly without performance degradation and at reduced compute.

In this configuration, we address a dental application. Specifically, we receive a stream of observations, i.e. frames, from a dental microscope. Dental microscopes are very important in any modern dental clinic. Their integration increases the quality of patient care and improves ergonomics for dentists. Despite their high magnification, dental microscopes are constrained by limited depth of field, leading to blurred images of moving dental instruments when focusing on intricate dental

structures. Moreover, the ability to identify the precise location of these instruments within videos holds significant potential for advancing medical video analytics. This Chapter explores the application of automated video deblurring and instrument segmentation in dental scenes.

We propose a deep multi-task learning architecture to explore simultaneous dental video deblurring and instrument segmentation. We posit that these two tasks are inherently interconnected, sharing common features tied to the motion dynamics within videos. Since spatially-variant blur kernels are formulated on the basis of moving objects, we adopt filter-adaptive kernels [148], proposed for motion deblurring, to mine the inter-frame spatio-temporal object boundaries. Consequently, we demonstrate the benefits of utilizing deblurring spatio-temporal features to enhance dental instrument segmentation. This approach not only outperforms state-of-the-art deblurring methods in terms of deblurring quality but also maintains robust segmentation performance, all while operating efficiently in terms of computational resources.

4.2 Method

We address the problem of dental instrument segmentation and deblurring from a multi-task learning perspective. Again, we assume that a camera streams frames at each time step t. We follow the notation of Chapter 3, and Equation 3.4. Let R_t and B_t be the full-resolution, sharp image and its blurry counterpart, respectively such that:

$$B_{x,t} = \langle (R_{x,t})^{-} k_{x,t} \rangle, \tag{4.1}$$

where $(R_{x,t})^{-}$ refers to a window of size K centered around pixel x in the image R_t, $k_{x,t}$ denotes the spatially variant blur kernel and $B_{x,t}$ is the resultant window on the blurry image. Estimating a blur-free frame $\hat{R}_t$ requires computing the inverse, that is an inherently ill-posed problem. Neural networks have proved successful in image and video deblurring due to implicit blur kernel approximations that bypass solving for the inverse of Equation 4.1. Simultaneously, the dental instrument mask is inferred from each blurry input frame. Likewise, we attempt to learn a non-linear function of parameters θ comprising $\mathcal{E}, \mathcal{D}_{deb}$ and $\mathcal{D}_{seg}$. The deblurring and segmentation processes are defined as composite functions of $\mathcal{E}$ and $\mathcal{D}_{deb}$ or $\mathcal{D}_{seg}$, i.e.

$$Deb = \mathcal{D}_{deb} \circ \mathcal{E} \tag{4.2}$$

$$Seg = \mathcal{D}_{seg} \circ \mathcal{E} \tag{4.3}$$

In our multi-task setting, $\mathcal{E}$ is shared among both tasks, where $\mathcal{D}_{deb}, \mathcal{D}_{seg}$ are task-specific.

To complement the input with informative cues from the past, in this work

we further employ the previous blurry and deblurred frames, as well as the previous segmentation output, i.e.

$$\hat{R}_t = Deb(\mathcal{CAT}_{10}(B_t, B_{t-1}, \hat{R}_{t-1}, \hat{M}_{t-1})) \tag{4.4}$$

$$\hat{M}_t = Seg(\mathcal{CAT}_{10}(B_t, B_{t-1}, \hat{R}_{t-1}, \hat{M}_{t-1})) \tag{4.5}$$

where $\hat{R}_t$ and $\hat{M}_t$ denote the estimated deblurred frame and segmentation mask at the t-th time step, respectively.

4.2.1 Architecture

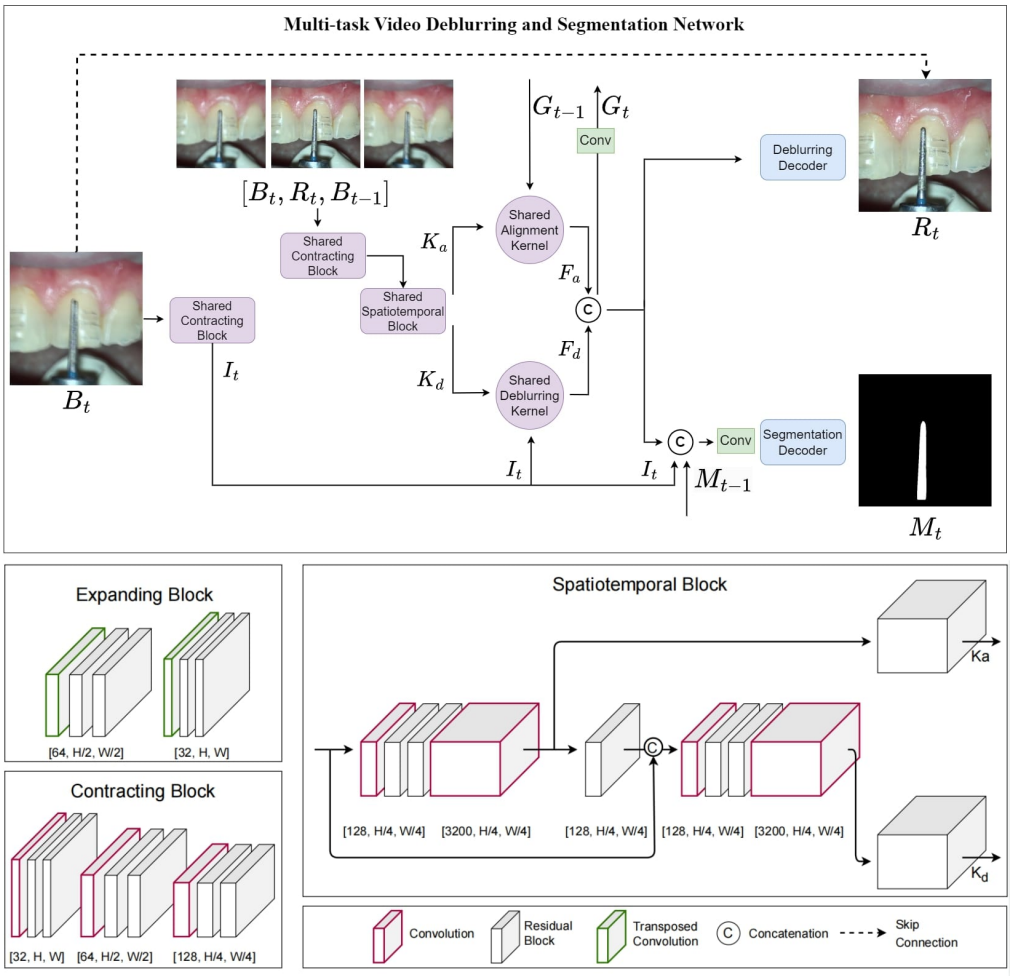


Figure 4.1: Overview of the proposed method. The top figure illustrates a higher level scheme whereas the bottom one contains depicts the architectural details.

In our proposed network, illustrated in Figure 4.1, we consider video streams, consisting of frames denoted as B_t, as the input to a contracting block resembling the U-Net architecture [73]. Within this block, two residual blocks [55, 148] follow

each strided convolution, resulting in output features denoted as I_t. Simultaneously, the current frame B_t is concatenated with the blurry and restored frames from the previous time step, resulting in $\mathcal{CAT}_9(B_t, \hat{R_{t-1}}, B_{t-1})$. This concatenated input undergoes a second contracting block with an identical structure, generating features termed as T_t. These features, denoted as T_t, are then processed by a spatiotemporal block to learn two essential components: alignment and deblurring kernels, represented as K_a and K_d, respectively. Specifically, K_a is responsible for aligning features from the previous time step, G_{t-1}, in a recurrent manner, while K_d focuses on deblurring the current frame on the feature level (I_t). Convolution operations using these kernels, applied to G_{t-1} and I_t, produce the transformed features F_a and F_d, respectively. These transformed features are further concatenated and passed to task-specific decoders. Finally, the concatenated features undergo a 3×3 convolution operation to create the memory state G_t.

The shared features are then fed into the decoders to accomplish the two primary objectives: restoring the blurry frame and predicting the segmentation mask. This is achieved through two U-Net-like expanding blocks, i.e. decoders D_{deb} and D_{seg}, where upsampling is performed using transposed convolutions, each followed by two residual blocks. Consequently, the restored, blur-free frame is obtained via:

$$\hat{R}_t = \mathcal{D}_{deb}(\mathcal{C}_{128,3\times3}(\mathcal{CAT}_{256}(F_a, F_d))), \tag{4.6}$$

whereas, for the segmentation branch we incorporate I_t and the mask predicted at the previous time frame, by stacking the channels as $(F_a, F_d, I_t, \hat{M}_{t-1})$ and transforming them by one 3×3 convolution, from 385 $(3 \times 128 + 1)$ to 128 feature maps. Likewise, the dental instrument at t is segmented as follows:

$$\hat{M_t} = \mathcal{D}_{seg}(\mathcal{C}_{128,3\times3}(\mathcal{CAT}_{385}(F_a, F_d, I_t, \hat{M}_{t-1}))), \tag{4.7}$$

where $\hat{M_t}$ is the two-channel (instrument vs background) pixel-wise segmentation scores followed by a softmax to transform the latter into probabilities.

4.2.2 Loss Function

To train the proposed network we resort to a linear combination of the task-specific loss functions. For video deblurring we employ the pixel-wise L2 loss, dubbed content loss, further augmented by a perceptual regularization term [138] that forces the learnt features to be close to the VGG-19 [3] ones, i.e.

$$\mathcal{L}_{content} = \frac{1}{CHW}||\hat{R} - R||^2 \tag{4.8}$$

and

$$\mathcal{L}_{perceptual} = \frac{1}{C_\phi H_\phi W_\phi}||\phi(\hat{R}) - \phi(R)||^2 \tag{4.9}$$

where C, H, W refer to the image channel, height and width dimensions respectively, ϕ denotes the VGG features and C_ϕ, H_ϕ, W_ϕ denote the corresponding feature dimensions.

For the semantic segmentation task, we utilize the binary pixel-wise cross-entropy loss function, i.e.

$$\mathcal{L}_{seg} = -\frac{\sum_i^{HW} M_i log(\hat{p}_i) + (1 - M_i) log(1 - \hat{p}_i)}{HW}, \tag{4.10}$$

where $\hat{p}_i$ denotes the probability that pixel i is predicted as part of the dental instrument segment and M_i the actual label. Therefore, the final loss function is defined as

$$\mathcal{L} = \mathcal{L}_{content} + \lambda_1 \mathcal{L}_{perceptual} + \lambda_2 \mathcal{L}_{seg} \tag{4.11}$$

where $\lambda_1 \in \mathbb{R}$, $\lambda_2 \in \mathbb{R}$ are weighting scalar values.

4.3 Experiments

4.3.1 Dataset

The dataset utilized in this research comprises actual dental treatment sessions recorded under a microscope. Initially, we amassed a collection of 12 FullHD video clips, capturing eight distinct dental treatment sessions. Subsequently, meticulous manual scrutiny of the intraoral snippets led to the removal of frames that were either blurred or devoid of informative content (e.g., instances of inactivity during the intraoral procedure). As a result of this curation process, the refined dataset consists of 67 video snippets, encompassing a total of 50241 frames.

For the purpose of labeling tooltip masks, a team of experts conducted manual annotations. To ensure diversity in the scenes and variations in instrument appearances, we uniformly sampled frames across the sequences. This effort resulted in 1,691 frames that were meticulously annotated with polygonal masks. To establish consistent segmentation labels throughout the sequences, we employed the MaskR-CNN [149] algorithm to interpolate these annotations. In our application, we treat different tooltip instances as a single entity.

To create realistic blur kernels, accommodating independent motion trajectories of various objects, we adopted a technique from [61]. This method involved the use of high-fps cameras to record videos, which were subsequently temporally averaged to generate blur. Given the constraint of our microscopic dataset, which was recorded at 29 fps, we employed each FullHD frame to generate multiple cropped (400×400) frames. The process entailed tracking the dental tooltip between frames $t-1$ and t, generating a path connecting the tooltip coordinates at these consecutive frames, and then moving a fixed-sized patch along that path. Each pixel coordinate along this path contributed to a patch. Finally, we averaged every 30 frames to cre-

ate the blurred frame and selected the middle frame within each window as the sharp frame. This workflow resulted in 67 sequences, each consisting of 97677 triplets of frames: B_i (blurred), I_i (sharp), and M_i (mask). The dataset was subsequently partitioned into training and test sets, comprising 56 sequences (83,390 frames) and 11 sequences (14,287 frames), respectively.

4.3.2 Setup

- Comparison with State-of-the-Art Models: We compared against state-of-the-art models, including *STFAN* [50] for deblurring and DLV3+ [90] with a *ResNet50* encoder for tooltip segmentation.
- Ablation Study: To assess the effectiveness of our multi-task learning setup, we compared our network, *MTL-AD-MPGT*, with *MTL-AD-MPP* to investigate the optimal mask propagation policy. Furthermore, we compare its predecessor, *MTL-AD*, with *MTL-A* and *MTL-V* to investigate the optimal, multi-task sharing pattern.

All experiments were conducted using an Nvidia Quadro RTX 5000 GPU and implemented in PyTorch [18]. We utilized the Adam optimizer [139] with a learning rate of 1e-4 and a batch size of 4 for all multi-task learning (MTL) networks. The training process for MTL networks consisted of 40 epochs, with a multiplicative learning rate decay factor of 0.5 applied every 10 epochs. Data augmentations incorporated random horizontal and vertical flipping, color jittering, and additive white Gaussian noise sampled with a variance of 0.14 to enhance the robustness of the models.

Regarding the linear scalarization of loss weights, we set λ_1 and λ_2 to 0.01 and 0.001, respectively. This choice was made to ensure that the contributions of the respective losses during training were balanced and had similar magnitudes. For the deblurring task, we followed the configurations provided by the authors for training the *STFAN* model. In the case of segmentation, we trained the *DeepLabV3+* [149] model with a *ResNet50* [55] backbone. The optimization process for segmentation was consistent with that of the MTL networks, including a learning rate decay by a factor of 0.1 in the 10th epoch. Training for segmentation also spanned 20 epochs and employed the same data augmentations and batch size as in the MTL experiments.

4.3.3 Results

Deblurring and segmentation performance are evaluated in PSNR and IoU, respectively, and runtimes are measured in frames per second (fps). Deblurring performance is consistent over all setups and in line with or higher than its single-task counterpart in all experiments (see Table 4.1). Best IoU performance on instrument segmentation is 87.1%, achieved by *DeepLabV3+* which introduces, however, expensive computations. Our *MTL-AD-MPGT* version achieves an IoU of 81.5%

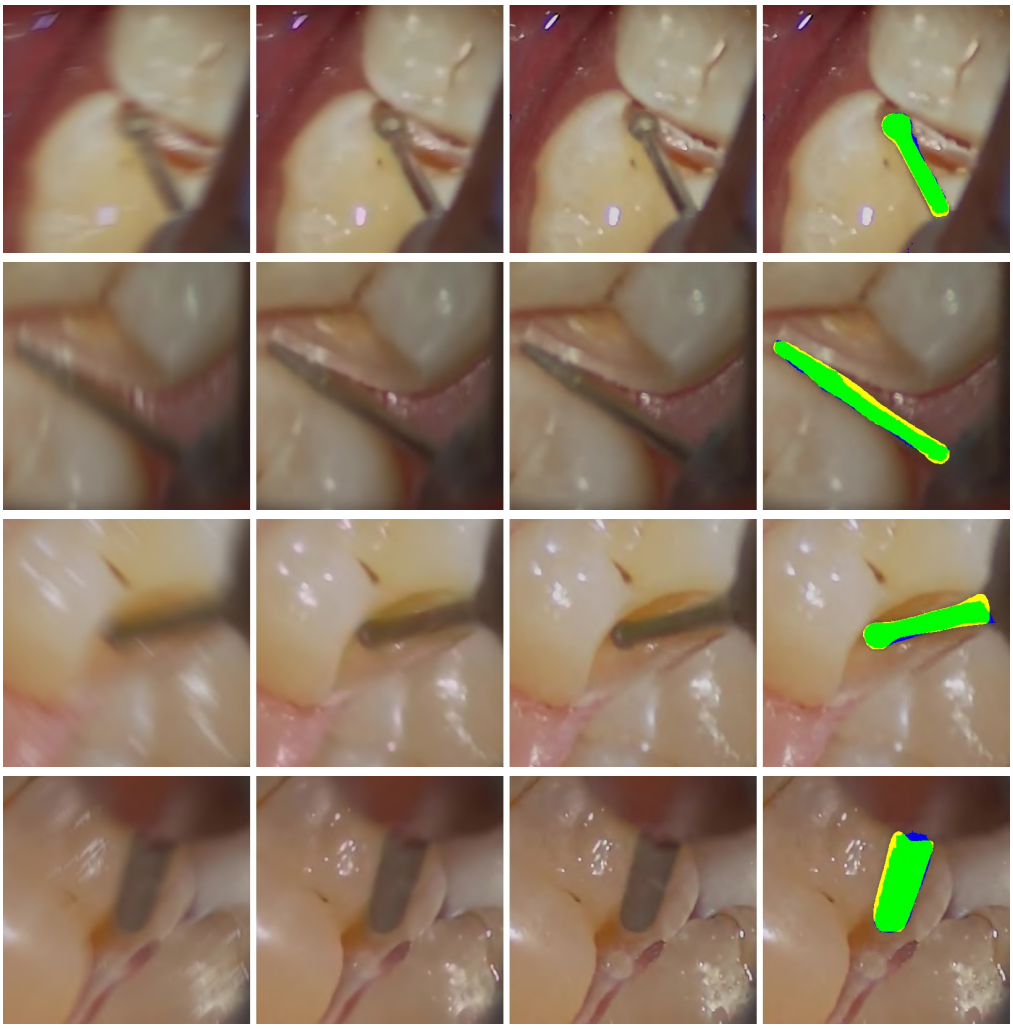


Figure 4.2: Qualitative results of the proposed method. From left to right: Input blurry frame (B), deblurred output frame ($\hat{R}$), GT sharp frame (R) and GT with overlaid mask ($\hat{M}$). Green, yellow and blue pixels correspond to TP, FP, FN, respectively. Best-viewed when zoomed in.

Table 4.1: Results of the proposed method compared to own baselines and state-of-the-art single-task solutions on the test set.

Architecture	**PSNR** ↑	**IoU** ↑	**FPS** ↑	**#P(M)** ↓
STL Deblurring STFAN [148]	38.91	-	31.2	5.4
STL Segmentation DLV3 [61]	-	87.09	32.2	26.7
STFAN + DLV3	38.91	**87.09**	16.1	32.1
MTL-V	39.06	68.14	21.3	6.8
MTL-A	38.90	75.66	21.3	7.2
MTL-AD	38.87	76.27	20.8	7.6
MTL-AD-MPP	38.90	76.33	20.8	8.2
MTL-AD-MPGT	**39.14**	81.46	**20.8**	**8.2**

improving performance over the *MTL-V* baseline by 13.3%. Simultaneously, *MTL-AD-MPGT* runs at 21 fps whereas the combined STL models run at 16 fps for the same workload. The qualitative results of the proposed method are illustrated in Figure 4.2. Clearly, the proposed method produces visually pleasant results restoring fine-grained image details while successfully segmenting the dental instrument.

4.4 Discussion and Conclusion

In this Chapter, we leverage the representational capacity of spatiotemporal features to address microscopic video deblurring and dental instrument segmentation in a multi-task learning configuration. We recurrently utilize the previous frame blur-free and mask model estimations as a guidance to predict the current ones. Likewise, the proposed method achieves higher PSNR performance than its single-task counterpart, yields a reasonably high IoU score for dental instrument segmentation and runs at 21 fps compared to the 16 fps of the combined single-task solutions. The multi-task network further achieves a $\times 4$ reduction in parameters compared to the system of the single-taskers, facilitating the deployment of networks on edge devices.

To conclude, the experiments conducted here reaffirm that low- and mid-level visual scene enhancement and understanding tasks such as the likes of video deblurring and segmentation are capable of being efficiently combined under an effective architecture. The results obtained in this Chapter allowed to experimentally prove the first hypothesis of this thesis. The proposed method and some results presented in this chapter were published in [150].

Chapter 5

Generalizing Diverse Vision Tasks with a Multi-output, Multi-scale, Multi-task Architecture

5.1 Introduction

In this Chapter, we investigate the second hypothesis of this dissertation and tackle a comprehensive range of video tasks spanning from low- to mid-level. Building upon the insights gained in our previous Chapters, we unify color mapping [151], denoising, and deblurring [62, 48, 142] into a single dense prediction task. Additionally, we incorporate homography estimation [81] and teeth segmentation [90, 89] as auxiliary tasks. These tasks are intricately interconnected; for example, aligning video frames aids in the process of deblurring and denoising [48, 50, 142, 62], while denoising and deblurring unveil image features that facilitate motion estimation [81] and further confirmed in the experimental offset visualization of Chapter 3. Moreover, semantic segmentation contributes positively to video deblurring, as demonstrated in Chapter 4, and Le et al. [81] highlight its contribution to homography estimation too. This Chapter represents the first study to jointly address color mapping, denoising, deblurring, motion estimation, and segmentation in a unified framework. The experimental part of this Chapter focuses again on RGB video scenes, addressing this time a medical application for multi-task video enhancement in dental videos.

The field of computer-aided dental intervention and macro-visualization, as discussed in previous works [152, 153, 154], presents a context where the tasks mentioned earlier find extensive applications. In modern dental practice, dentists utilize a range of tools to enhance their view of teeth, aiming to reduce the time required for procedures while improving their quality [155]. Maintaining a close and un-

interrupted view of the tooth being operated on is crucial for performing dental bur maneuvers effectively and safely, particularly when removing caries, in order to minimize the risk of exposing the pulp tissue to infection. To achieve this, a microcamera, attached securely to a dental handpiece near the dental bur, allows dentists to closely and continuously inspect the tooth during drilling through a display. However, the need for miniaturization of vision sensors and optics introduces imaging artifacts. Macro-view magnifies the slight movements of the bur, resulting in significant image displacement. Continuous camera shakes can lead to eye fatigue and image blurring. Additionally, handpiece vibrations, rapid changes in lighting conditions, and the presence of splashing water and saliva further complicate the imaging of intra-oral scenes. This study is the first to address the effectively compromised quality of videos of phantom scenes with an algorithmic solution to integrate cost-effective microcameras into digital dental workflows.

In this Chapter, we propose *MOST-NET* (multi-output, multi-scale, multi-task), a network designed to model and harness task interactions across various scale levels within the encoder and decoder. *MOST-NET* is a new multi-task, decoder-focused architecture [156] for video processing. The proposed network has multiple heads at each scale level. Provided that task-specific outputs amend themselves to scaling, the network propagates the outputs bottom-up, from the lowest to the highest scale level. It thus enables task synergy by loop-like modeling of task interactions in the encoder and decoder across scales. Different than state-of-the-art multi-task networks such as MTI-Net [22], that propagates the task features in scale-specific distillation modules across scales to the encoder, our network simultaneously propagates task outputs to the encoder and to the task heads in the decoder. Furthermore, all previous works make dense task prediction in static images while we our network is the only method explicitly applicable to video scenes.

5.2 Method

We address the video enhancement tasks in dental interventions. In this specific scenario, we set $T = 3$, and $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{O}_3$ are used to represent the outputs for video restoration, segmentation, and homography estimation, respectively. The video stream produces observations $\{B_{t-p}\}_{p=0}^{P}$, where t serves as the time index, and $P > 0$ represents the number of preceding frames. The objective here is to predict a clean frame, generate a binary teeth segmentation mask, and estimate inter-frame motion using a homography matrix. This information is encapsulated within the triplet denoted as $\mathcal{O}_{1,1}^{3,S} = \{R_t^s,\, M_t^s,\, H_{t-1\to t}^s,\}_{s=1}^{S}$.

The degradation problem is similar to that of Section 3.2. Let x correspond to pixel location. Given per-pixel blur kernels $k_{x,t}$ of size K, the degraded image at $s = 1$ is generated as:

$$\forall_x\ \forall_t\ B_{x,t} = \sigma_n \langle (R_{x,t})^{-} k_{x,t} \rangle + \eta_n, \tag{5.1}$$

Here, η_n represents additive noise, and σ_n stands for signal-dependent noise. Additionally, $(R_{x,t})^-$ refers to a window of size K centered around pixel x in the image R_t^1. Next, we make an assumption that multiple independently moving objects are present within the scenes we are considering. However, our primary task is to estimate the motion related to the specific object of interest (in this case, teeth), which is confined to the region indicated by non-zero values in the mask M:

$$\forall_t \ \forall_x \qquad x_t = H_t x_{t-1} \quad s.t. \quad M_t(x) = 1 \tag{5.2}$$

5.2.1 Architecture

The architecture, presented in Figure 5.1, follows a U-shaped [73] structure for feature extraction. These features are aligned with the previous frame's features through homographies from preceding scales. Thereafter, the per-scale encoder features are used to estimate the dense outputs and the homograpy, at each scale, in a bottoms-up manner. This proposed bottoms-up output propagation facilitates the feeding of lower-scale outputs as inputs to higher scales, across tasks. The proposed multi-task, multi-output, multi-scale (*MOST*) design fosters interaction to enhance overall multi-task performance.

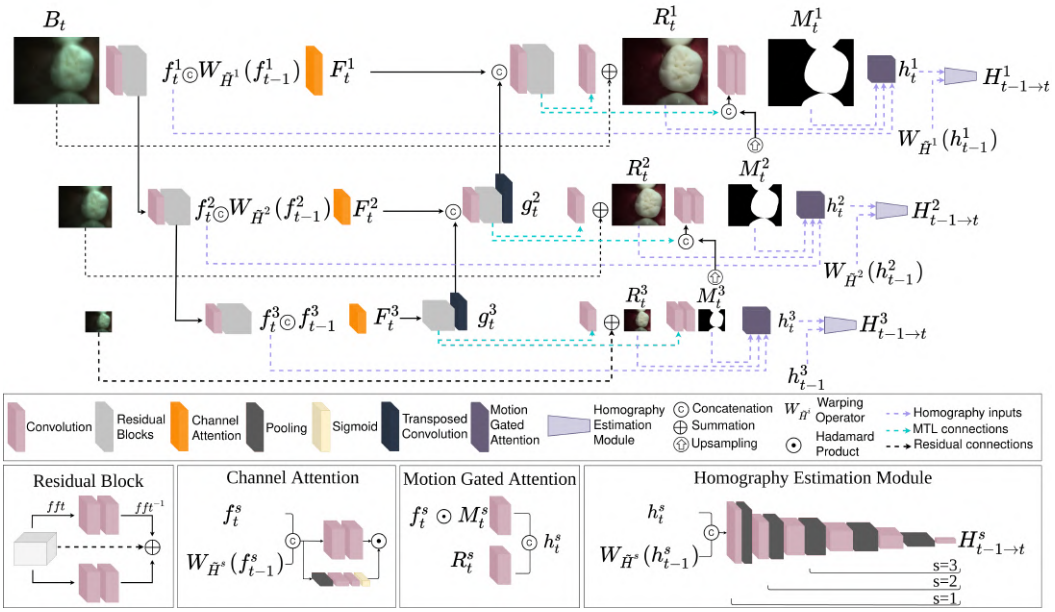


Figure 5.1: Our *MOST-NET* instantiation addresses three tasks for video enhancement: video restoration, teeth segmentation, and homography estimation.

Encoders: Feature Extraction

At each time step, *MOST-NET* independently extracts features f_{t-1}^s and f_t^s from two input frames, B_{t-1} and B_t, at three different scales. To achieve downsampling in a U-shaped manner [73], features are obtained through 3×3 convolutions with strides of $1, 2, 2$ for $s = 1, 2, 3$, followed by ReLU activations and augmented by

five residual blocks [49] at each scale. These residual connections are enhanced with an additional branch of convolutions in the Fast Fourier domain [157] as shown in Figure 5.1. The output channel dimension for features f_t^s is set to 2^{s+4}.

Encoders: Feature Alignment

At each scale, features f_t^s and $W_{\tilde{H}_s}(f_{t-1}^s)$ are concatenated, and a channel attention mechanism is applied following the approach described in [48], resulting in F_t^s. To warp encoder features from the previous time step, *MOST-NET* employs homography outputs from lower scales, denoted as $W_{\tilde{H}}(f_{t-1}^s)$. Here, $\tilde{H}^s$ represents an upscaled version of H^{s+1} for higher scales and the identity matrix for $s = 3$.

Decoders: Dense Outputs

The attended encoder features F_t^s are passed onto the expanding blocks scale-wisely via the skipping connections. At the lower scale ($s = 3$), the attended features F_t^3 are directly passed on a stack of two residual blocks with 128 output channels. Thereafter, transposed convolutions with strides of 2 are used twice to recover the resolution scale. Moving to higher scales ($s < 3$), the features F_t^s are initially combined with the upsampled decoder features and convolved using 3×3 kernels to reduce the number of channels by half. Following this, they are passed through two residual blocks, each with 64 and 32 output channels, respectively. The outputs of these residual blocks form scale-specific shared backbones. Lightweight, task-specific branches follow then to estimate the dense outputs. Specifically, one 3×3 convolution is used to estimate R_t^s, and two 3×3 convolutions, separated by ReLU activations, produce M_t^s at each scale. The architecture of *MOST-NET*, as illustrated in Figure 5.1, facilitates the refinement of lower scale segmentations through upsampling and their input to the task-specific branches at higher scales.

Decoders: Homography Outputs

At each scale, the homography estimation modules calculate four offsets, which have a one-to-one correspondence with homographies through the Direct Linear Transformation (DLT), as demonstrated in [78, 81]. The motion gated attention modules then perform element-wise multiplication between the features f_t^s and the segmentations M_t^s, which helps filter out context that is not relevant to the motion of the teeth. Subsequently, a 3×3 convolution is used to reduce the channel dimensionality by half, while a second convolution extracts features from the restored output R_t^s. The combination of these two streams results in the formation of features h_t^s. At each scale, these features, denoted as h_t^s and $W_{\tilde{H}^s}(h_{t-1}^s)$, are utilized to predict the offsets using shallow downstream networks. Predicted offsets at lower scales are then converted back to homographies and cascaded in a bottom-up fashion, following a similar approach to [81]. In a manner similar to [78], we employ blocks of 3×3 convolutions, coupled with ReLU activations, batch normalization, and max-pooling, to reduce the spatial dimensions of the features. Just before the regression layer, a dropout of 0.2 is applied. For $s = 1$, the convolution output channels are set to 64, 128, 256, 256, and 256. For $s = 2$ and $s = 3$, the network depth is cropped from the second and third layers onwards, respectively.

5.2.2 Loss Function

In our context, the multi-task dataset is denoted as $\mathcal{D} = \{\{B\}_j, \{\mathcal{O}_i^s\}_j\}_{i,s,j=1}^{T,S,N}$, where $\{\mathcal{O}_i^s\}_j$ is a label related to task i at scale s for the j-th training sample $\{B\}_j$, while N denotes number of samples in training data. The optimal set of parameters θ for *MOST-NET* is derived by minimizing the objective below:

$$\mathcal{L}(\theta) = \sum_{i}^{T} \sum_{s}^{S} \lambda_i \mathcal{L}_i \left(\mathcal{O}_i^s, \hat{\mathcal{O}}_i^s(\theta) \right), \tag{5.3}$$

where λ_i is a scalar weighting value, $\hat{\mathcal{O}}_i^s(\theta)$ is an estimate of $\mathcal{O}_i^s$ for j-th sample in $\mathcal{D}$, and $\mathcal{L}_i$ is a distance measure. Please note that we are ommitting the subscript j for notational brevity; the loss values across scales and tasks are summed on the batch-level. In this Chapter, we adopt the Charbonnier loss [158] as $\mathcal{L}_1$, the binary cross-entropy [159] as $\mathcal{L}_2$ and the Mean Average Corner Error (MACE) [78] as $\mathcal{L}_3$.

5.3 Experiments

5.3.1 Dataset

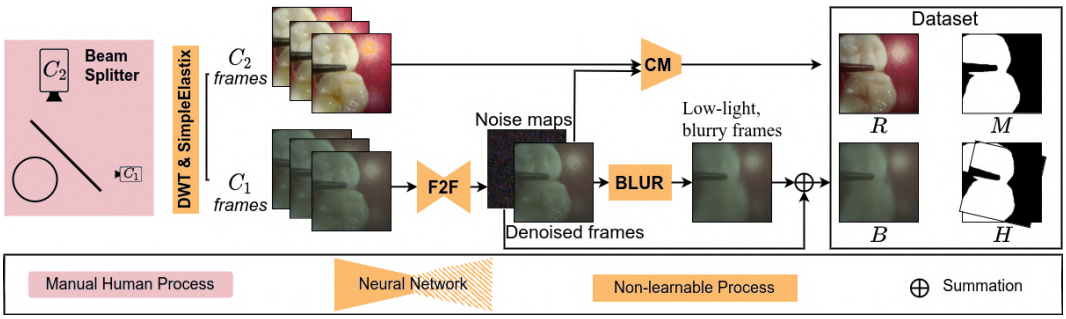


Figure 5.2: A flowchart of dataset preparation.

We present the acquisition and generation process of the *Vident-lab* dataset, denoted as $\mathcal{D}$, which includes frames B and labels R, M, and H for the purposes of training, validation, and testing (refer to Table 5.1). These labels are generated at full resolution, as visualized in Figure 5.2. For the lower scale labels $\mathcal{O}_i^s$, we derive them with downsampling (bilinear interpolation) for R and M, as well as downscaling for H. The dataset is publicly available [29].

Data acquisition

In our setup, a miniaturized camera labeled as C_1 exhibits lower image quality compared to the intraoral cameras designated as C_2, benefiting from larger sensors and advanced optics. Our objective revolves around training C_1 to image the dental scene with the same level of quality as achieved by C_2. Both cameras are firmly coupled through a 50/50 beam splitter, allowing them to simultaneously record videos of the identical dental scene. We employ Dynamic Time Warping (DTW) to

synchronize the videos, followed by the use of *SimpleElastix* [160] for registering the corresponding 320 × 416 frames.

Data	Train	Validation	Test
Videos	300	29	80
Frames	60K	5.6K	15.5K
Segm (H)	300	116	320
Segm (N)	59.7K	5.5K	15.2K

Table 5.1: Dataset summary ($K = \times 10^3$), (H,N) human- and network-labelled teeth masks.

Noise, blur, colorization

We adopt frame-to-frame (F2F) training [161] on frames captured by camera C_1. We then use the trained image denoiser to produce *denoised frames* and their corresponding *noise maps*. To introduce realistic blur, we temporally interpolate the denoised frames eight times, followed by averaging them over a temporal window spanning 17 frames. The noise maps are subsequently added to the blurred frames, forming the *input video frames* denoted as B. However, achieving perfect frame registration between two distinct modalities like C_1 and C_2 presents a challenge. To overcome this, we employ colorization techniques to transform the frames from C_1 to match those from C_2, creating the ground truth *output video frames* represented by R. Similar to the approach used in [151], we train a Color Mapping (CM) network to predict parameters of 3D functions, facilitating the mapping of colors from dental scenes captured by C_2 to those captured by C_1. This method allows us to bypass local registration errors, ensuring an exact pixel-to-pixel spatial correspondence between frames B and R.

Segmentation Masks and Homographies

We manually annotate a single frame R containing natural teeth within the phantom scenes from each training video. In the validation and test video sets, we annotate four frames in each video. Following the approach outlined in [162], we fine-tune an HRNet48 [163] pretrained on ImageNet using our annotations to automatically segment teeth in the remaining frames across all three splits. We compute optical flows between consecutive clean frames using RAFT [84]. These motion fields are then cropped using teeth masks M_t to eliminate other moving objects such as dental bur or the suction tube, as our primary interest is to estimate the motion with respect to the teeth. Subsequently, we fit a partial affine homography H using RANSAC to the segmented motion field.

5.3.2 Setup

We perform multiple experiments to showcase a) the effectiveness of the multi-scale nature of *MOST-NET* b) the synergy of the proposed architecture and c) its performance compared to multiple state-of-the-art single task baselines.

- We assess performance gains across scale levels of *MOST-NET*. To this end, we upsample all outputs at lower scales to original scale and compare them with ground truth.
- We reconfigure *MOST-NET* by removing different branches to showcase that the proposed design is the most optimal.
- We compare *MOST-NET* with single task state-of-the-art methods for restoration, homography estimation and segmentation.

We train, validate, and test all methods on our dataset, as reported in Table 5.1. In all *MOST-NET* training runs, we set $\lambda_1, \lambda_2, \lambda_3$ to 2×10^{-4}, 5×10^{-5} and 1 for balancing tasks in Equation 5.3. Training for all methods is performed with a batch size of 16 and employs the Adam optimizer with an initial learning rate of $1e-4$, which is decayed to $1e-6$ using cosine annealing. Augmentation of the training frames includes horizontal and vertical flips with a probability of 0.5, random channel perturbations, and color jittering, following the methodology introduced in [50]. All experiments are carried out using PyTorch 1.10, and the inference speed is reported in frames per second (FPS) on an NVidia RTX 5000 GPU.

5.3.3 Results

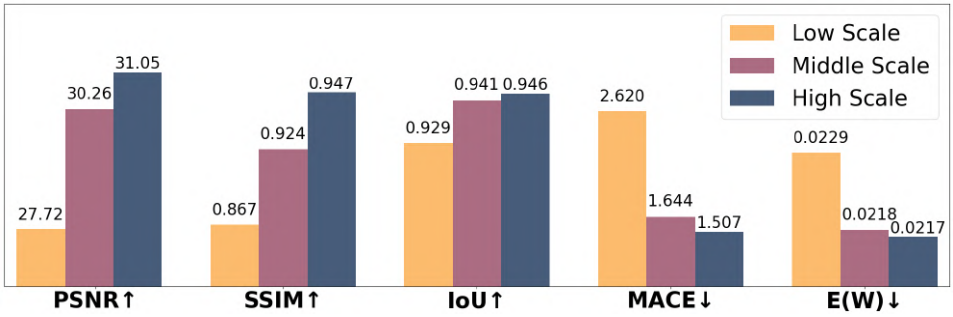


Figure 5.3: MOST-Net performance improves with output upscaling.

In Figure 5.3, we evaluate the performance improvements achieved across different scale levels of *MOST-NET*. Our observations indicate that the performance of *MOST-NET* exhibits improvement through the propagation of task output across scales, as reflected in all measured metrics. To further analyze the impact of various components of MOST-NET, we conduct an ablation study presented in Table 5.2. The architectural modifications for the ablation study include: (i) NS — omission of segmentation as an auxiliary task, (ii) NE — exclusion of the connection between encoder features f_t^s and the Motion Gated Attention module, (iii) NW — elimination of the warping of previous encoder features f_{t-1}^s, and (iv) NMO — disabling multi-outputs at lower scales i.e. $s > 1$, resulting in no task interactions between scales.

The ablation results reveal that all configurations lead to a decrease of more

than 0.5dB in PSNR and an increase in temporal consistency error E(W). Particularly, the inclusion of the segmentation task and temporal alignment proves most beneficial for the video restoration task. Specifically for NE, the decrease is 0.7dB suggesting again better frame feature representations, when frames are aligned on the feature level. The necessity of alignment occurs once more throughout this dissertation; it suggests that aligning frames is beneficial, here, by highlighting the interaction between the restoration and homography tasks. The absence of multi-task interactions across scales results in a MACE error increase of more than 0.6. The NE ablation demonstrates a slight improvement in MACE, albeit with a significant drop in PSNR. Interestingly, the IoU remains relatively unaffected by the ablations, suggesting potential for enhancing task interactions to assist in the segmentation task. Qualitative results are presented in Figure 5.4.

Architecture	**PSNR** ↑	**SSIM** ↑	**MACE** ↓	**IoU** ↑	**E(W)** ↓	**#P(M)** ↓	**FPS** ↑
MIMO-UNet [49]	26.66	0.916	-	-	0.0278	5.3	8.4
ESTRNN [48]	30.72	0.943	-	-	0.0229	2.3	68.5
MHN [81]	-	-	1.347	-	-	6.2	89.8
DeepLabv3+(DL) [90]	-	-	-	0.968	-	26.7	108.2
UNET++ [89]	-	-	-	0.969	-	50.0	38.9
ESTRNN+MHN+DL	30.72	0.943	**1.368**	**0.967**	0.0229	35.2	**28.6**
MOST-NET-NS	30.21	0.939	1.426	-	0.0223	9.7	19.0
MOST-NET-NE	30.22	0.941	1.423	0.946	0.0221	9.8	19.2
MOST-NET-NW	30.37	0.943	1.456	0.952	0.0221	9.8	19.3
MOST-NET-NMO	30.48	0.940	2.155	0.946	0.0227	8.5	19.1
MOST-NET	**31.05**	**0.947**	1.507	0.946	**0.0217**	**9.8**	19.3

Table 5.2: STL and MTL benchmarks (top panel) and *MOST-NET* (bottom panel). Best results of *MOST-NET* compared to ESTRNN+MHN+DL are in bold.

We present a comparison of *MOST-NET* with state-of-the-art single-task methods in the tasks of restoration, homography estimation, and binary segmentation, as summarized in Table 5.2. In the context of video restoration, *MOST-NET* surpasses the performance of the ESTRNN baseline [48] and the image restoration model MIMO-UNet [49] by more than 0.3dB and 4.3dB in PSNR, respectively. We attribute the lower PSNR performance of MIMO-UNet to its reliance on a single-frame input, impacting its colorization capabilities and leading to elevated temporal consistency error E(W) [164]. Additionally, ESTRNN exhibits observable flickering artifacts, as evident from its higher E(W) compared to *MOST-NET*. In homography estimation, *MOST-NET* marginally trails behind MHN [81] in terms of MACE error. However, our multi-tasking approach suggests potential for improvement in homography estimation. Notably, MHN achieves significantly lower MACE error when trained and tested on ground truth videos, emphasizing the necessity of the video restoration task to enhance homography estimation. Moving on to binary segmentation tasks, we benchmark *MOST-NET* against DeepLabv3+ [90] with ResNet50 encoder and UNET++ [89] for teeth segmentation using the intersection-over-union (IoU) criterion. *MOST-NET* achieves comparable results with significantly fewer parameters, maintaining near real-time efficiency despite addressing three tasks concurrently.

Lastly, as there is other such multi-task method available, we compare our multi-task *MOST-NET* with a forked pipeline of single-task methods—ESTRNN for video restoration, MHN for homography estimation, and DeepLabv3+ (DL) for segmentation. Despite running at 28.6 FPS, the forked pipeline requires 3.6 times more model parameters than *MOST-NET*. Moreover, *MOST-NET* outperforms the forked pipeline in PSNR, SSIM, and E(W) for video restoration, with comparable MACE error and IoU scores, while operating near real-time at either 19.3 FPS or **21.3 FPS** via TorchScript.

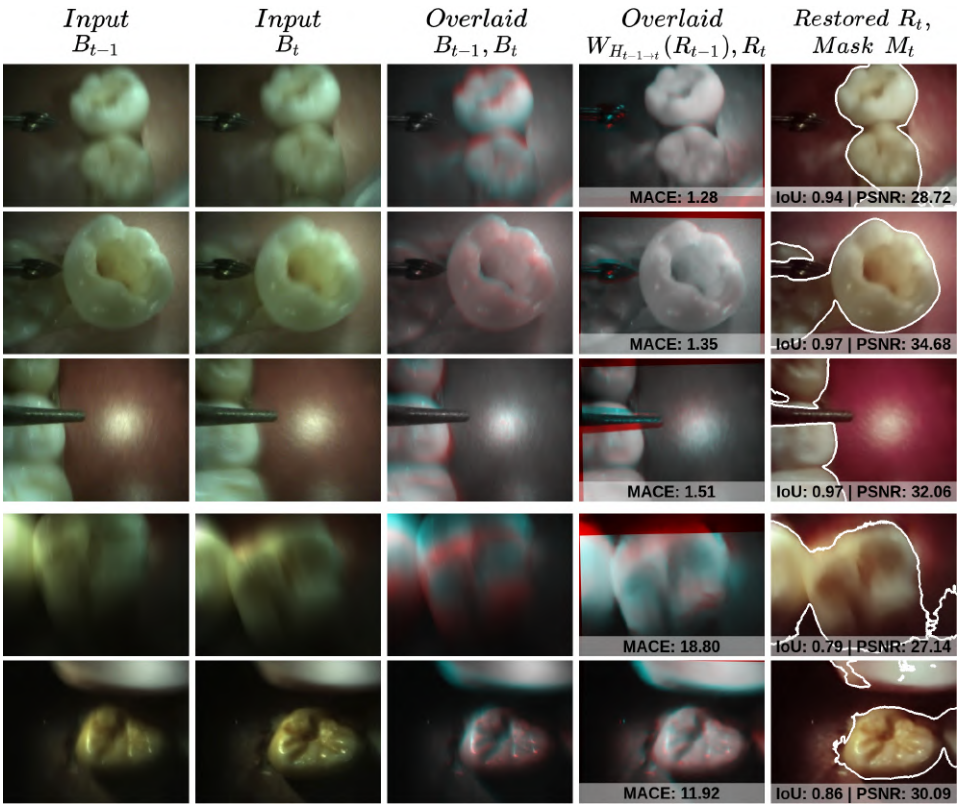


Figure 5.4: Our qualitative results of teeth-specific homography estimation (4th column) and full frame restoration and teeth segmentation (5th column). *MOST-NET* can denoise video frames and translate pale colors (first and second column) into vivid colors (5th column). Simultaneously, it can deblur and register frames wrt to teeth (4th column). In addition, despite blurry edges in the inputs, *MOST-NET* produces segmentation masks that align well with teeth contours (rows 1-3). Failure cases (bottom panel, 4-5th rows) stem from heavy blur (4th row, and tooth-like independently moving objects (5th row), such as suction devices.

5.4 Discussion and Conclusion

In this Chapter, we introduced *MOST-NET*, an novel deep neural network designed for comprehensive video processing by modeling task interactions across multiple scales. This novel architecture concurrently tackles a wide range of tasks spanning from low- and mid-level scene enhancement, encompassing deblurring, denois-

ing, and color mapping, to low- and mid-level understanding tasks such homography-based motion estimation or even teeth segmentation. Once more, this Chapter confirms the necessity for frame alignment and shows how homography estimation, as an auxiliary task, can improve performance on other video tasks. Our study highlights the practical applicability of *MOST-NET* within the context of computer-aided dental interventions. In this Chapter, we confirm our second hypothesis; one can accommodate diverse tasks in a multi-task architecture for RGB video scenes. Indeed, *MOST-NET* has lower parameter count than combined, state-of-the-art, single-task models and fast runtimes, at lower parameter count, despite yielding multiple task outputs per scale.

To facilitate research in the integration of visual scene enhancement and understanding tasks, we have openly shared the *Vident-lab* video dataset. This dataset features natural teeth within phantom scenes and can serve as a resource for training on a diverse set of tasks. The dataset accompanies the Katsaros et al. [165]. Portions of research from the respective paper are discussed and generalized in this Chapter.

The *MOST-NET* network has shown promise, but has its own limitations. Firstly, it is currently limited to working with RGB color data, and its effectiveness has only been demonstrated in macro-visualization settings. While expanding its capabilities to other domains is an exciting possibility, our current findings are constrained by the absence of relevant datasets. Secondly, while *MOST-NET* incorporates video denoising, deblurring, and color correction, these functionalities are combined and treated as a single task of visual scene enhancement, as detailed in Chapter 3. This approach, while effective, raises concerns about the scalability of the architecture when handling more decoders to produce separate outputs, which effectively increases the number of different gradient sets.

Chapter 6

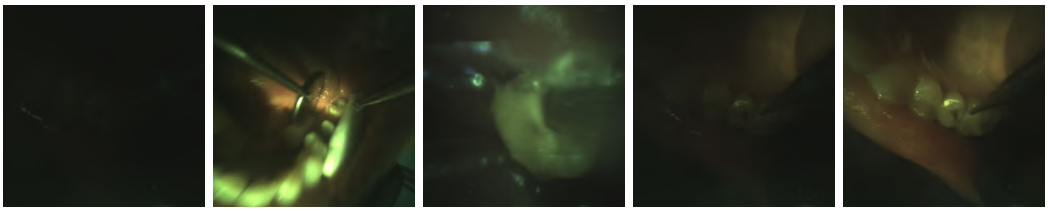
Neural Scale Search and Adaptive Task Balancing

6.1 Introduction

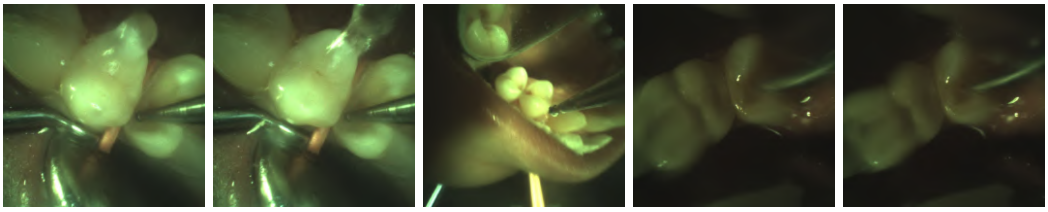
This Chapter revolves around the application discussed in Chapter 5, this time focusing on a new dataset, characterized by distinct challenges. The freshly introduced dataset stems from real intra-oral dental procedures conducted at the Medical University of Gdansk. This new dataset brings multiple challenges, such as reduced lighting conditions, heightened noise levels, and scenes with varying depths and teeth appearance that pose challenges to achieving optimal homographies and segmentation maps. While the pool of tasks remains the same, the number and severity of artifacts or challenges are escalated. The novel challenges are clearly illustrated in Figure 6.1. This dataset will also be made publicly available to facilitate relevant research.

To address the aforementioned issues, the preceding Chapter explored a multi-task, decoder-focused model [156] for multi-output, multi-scale, and multi-task video enhancement and understanding. This model, with multiple heads, predicts all multi-task outputs at each scale level. The network propagates the per-task outputs bottoms-up, from the lowest to the highest scale level. Likewise, it enables task interactions through a loop-like modeling of multi-task relationships in both the encoder and decoder across scales. Consequently, it allows for improved task interactions, refines predictions across scales, and offers insights into the contribution of each scale to the performance improvement of each task.

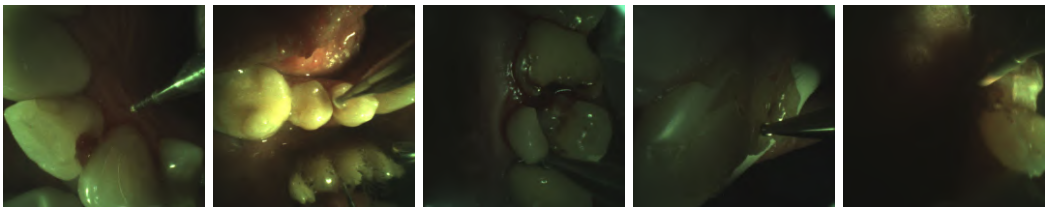
In this Chapter, our third hypothesis assumes that not all scales are equally crucial for all tasks, which coincides with the study of prevailing design choices in state-of-the-art networks. Tasks exhibit varying degrees of granularity requirements, with networks addressing low-level tasks allocating more computations at higher scales, while those designed for higher-level tasks prioritize lower scales. For instance, networks dedicated to tasks like deblurring [48, 49] and denoising [47, 46] focus their



(a) low light (b) motion blur (c) occluding water (d) light change: t (e) light change: $t+1$



(f) tooth deformation: t (g) tooth deformation: $t+1$ (h) teeth at different depths (i) tooth deformation: t (j) tooth deformation: $t+1$



(k) white gloves (l) teeth in mirror (m) blood (n) lack of context - close view (o) lack of context - missing teeth

Figure 6.1: Processing real dental videos in a multi-task setting poses significant challenges due to factors such as camera miniaturization and scene characteristics influenced by artifacts, parallax, non-rigidity, ambiguity, and texture scarcity. **Top row** examples highlight key restoration challenges: dark images (a), blur (b), water interference (c), and sudden light changes across frames (d, e). **Middle row** instances feature challenges in homography estimation: tooth deformations caused by water (f, g, i, j) and teeth at varying depths, leading to different motion planes (h). **Bottom row** image provide examples illustrating difficulties in teeth segmentation, including objects with colors resembling teeth (e.g., gloves or sponges) (k), mirrored teeth (l), blood on drilled regions (m), a lack of contextual cues for segmentation in close-up views (n), and scenarios with multiple missing teeth (o), all of which complicate the segmentation task.

computations closer to the original input image scale to preserve spatial information. Conversely, mid- or high-level tasks such as optical flow [166, 84], segmentation [90], or surgical tool pose estimation [167] allocate more computation at lower scales of the original image size to enhance reasoning about motion and spatial context.

Learning multiple tasks in a unified architecture requires rigid assumptions on a) the impact of scales on each individual task, b) the interactions between the tasks. The increase in performance across scales may be non-uniform and can vary across tasks. Additionally, the coexistence of signals from multiple tasks across various scales often leads to a significant number of gradient conflicts, thereby impeding overall performance, as noted by Yu et al. [35]. Our fourth hypothesis challenges the notion that all gradients are equally essential at every scale, as some gradients may introduce conflicts rather than contribute positively. To address this, we propose Neural Scale Search (NSS), a gradient-based approach that explores the optimal scales-for-tasks structure within the output space of the *MOST* model. NSS leverages Softmax-Gumbel continuous relaxation to navigate the discrete search space based on innerscale output differences. By quantifying the significance of task outputs across scales in terms of their contribution to performance improvement (or surrogate loss minimization as a proxy for each task's metric), NSS introduces *MOST-NSS++*, an architecture that is both sparser and more efficient than the original *MOST* network.

Even when equipped with a suitable multi-task architecture and appropriate linear scalarization of the multi-task loss weights, the optimization process of multi-task networks remains challenging. Here, we discuss the second part of our fourth hypothesis. We note that the initially balanced multi-task training achieved through carefully assigned loss weights i.e. linear scalarization, may become uneven at later stages, resulting in disparate learning paces. While existing approaches aim to address variations in multi-task gradient magnitudes and directions, they do not explicitly tackle the issue of uneven multi-task training and often require access to network gradients, incurring a linearly growing memory cost with the number of tasks. In our approach, designed with consideration for dataset size, we overcome the computationally expensive access to gradients by introducing Adaptive Task Balancing (ATB), a straightforward yet highly effective weighting scheme. ATB dynamically adjusts task weights during training, ensuring that tasks progress at comparable rates.

6.2 Method

In this Chapter, the problem formulation is identical to the one of Chapter 5, as illustrated in Equation 5.1. The multi-task convolutional network attempts to solve multiple tasks via three outputs,i.e. the restoration and segmentation images, as well as the homography relating two consecutive frames. In the next Subsection, we initially introduce a new MOST architecture, i.e. *MOST-NET++*. Thereafter, we

discuss our two final hypotheses, i.e. optimal resource allocation via an end-to-end, NAS-inspired approach to restructure the gradient flow and discover a sparser, yet more efficient architecture and mitigation of the diverse multi-task training paces.

6.2.1 Architecture

Here, we describe the architecture we will employ as a baseline to address the third and fourth hypothesis of this thesis. Specifically, we modify the *MOST-NET* architecture into *MOST-NET++*, a novel instantiation of the *MOST* model.

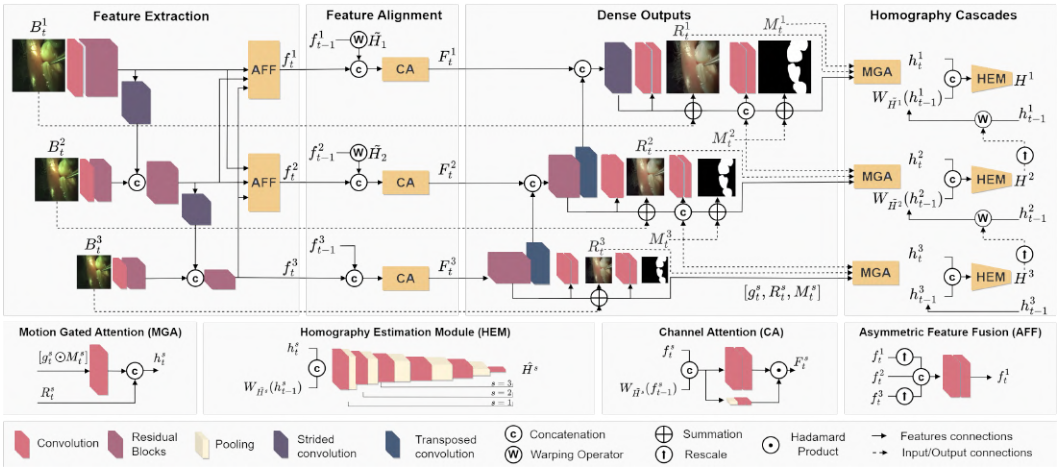


Figure 6.2: The proposed *MOST-NET++* achieves multi-scale feature exchange and alignment, at the encoder level, and bottom-up multi-task output interaction and refinement across scales, at the decoder level.

As illustrated in Figure 6.2, the encoder consists of a feature extractor that combines deep- and image-level features at each scale and a feature alignment module that aligns the previous frame to the current and fuses their information via channel attention. In contrast to Chapter 5, this architecture employs additionally Asymmetric Feature Fusion, image-level features, batch normalization across all residual blocks. To make our network more efficient for real-time applications we omit the Fourier transforms and convolutions and focus solely on the pixel domain.

Similarly to Chapter 5, the decoding part consists of the dense outputs, i.e. the decoders branch out scale-wise to produce one output for each task and scale. The scale-wise decoders are also shared with the homography estimation modules that estimate and refine the homography bottoms-up in a cascade. It is different to that segmentation maps are refined in a residual fashion across scales. Moreover, the task-specific segmentation heads are augmented with one additional convolutional layer. The details of the architecture are discussed in the next subsections.

Encoders: Feature Extraction

The cameera streams frames B_t at each time step t. *MOST-NET++* extracts features f^s_{t-1} and f^s_t independently from two input frames B_{t-1} and B_t at three different scales. To achieve U-shaped downsampling [73], deep features are obtained

through 3×3 convolutions with strides of $1, 2, 2$ for $s = 1, 2, 3$, followed by ReLU activations. Downsampling convolutions at lower scales (i.e., $s = 2, 3$) result in some loss of spatial information. In this study, we address this issue by first combining the deep features with image-level features, extracted from the downsampled image itself, following the approach of MIMO-UNet [49]. These concatenated, deep- and image-level, features are more descriptive representations of the image. Thereafter, they undergo processing through a sequential stack of 5 residual blocks at each scale. Second, we improve the residual output features through cross-scale information exchange, illustrated in Figure 6.2, specifically for $s = 1, 2$, utilizing Asymmetric Feature Fusion (AFF)[49]. Exchanging information across scales improves representations too, as it allows the higher scale to complement its features with features of a wider effective receptive field and vice-versa. The output channel dimension for features f_t^s is 2^{s+4}. Differing from Chapter 5, we completely exclude Fourier transforms and convolutions and instead employ plain residual blocks.

Encoders: Feature Alignment

At each scale, the features f_t^s and $W_{\tilde{H}_s}(f_{t-1}^s)$ undergo concatenation, followed by a channel attention mechanism [48]. This process selectively fuses the features into F_t^s. In the context of*MOST-NET++*, the homography outputs from lower scales are employed to warp encoder features from the preceding time step, represented as $W_{\tilde{H}}(f_{t-1}^s)$. In this expression, W denotes the warping operator, while $\tilde{H}^s$ represents an upscaled version of H^{s+1} for higher scales and the identity matrix for $s = 3$.

Decoders: Dense Outputs

Following the channel attention mechanism, the attended encoder features F_t^s are conveyed to the expanding blocks in a scale-specific manner through the use of skipping connections. At the lower scale ($s = 3$), the attended features F_t^3 directly feed into a stack of two residual blocks, each producing 128 output channels. Subsequently, two transposed convolutions with strides of 2 are applied to restore the resolution scale. For higher scales ($s < 3$), we concatenate F_t^s with the upsampled decoder features from the lower scale. This concatenated result is then convolved using 3×3 kernels to reduce the channel count by half. The output is then passed through two residual blocks, generating 64 and 32 output channels, respectively. These outputs from the residual blocks serve as scale-specific shared backbones, resulting in output features g_t^s.

Following this, lightweight task-specific branches are introduced to estimate dense outputs at each scale. In this study, the number of task-specific layers for estimating M_t^s and R_t^s at each scale is increased to a total of three 3x3 convolutions, interspersed with ReLU activations. Figure 6.2 visually illustrates how *MOST-NET++* facilitates the refinement of lower scale segmentation by upsampling and feeding them into the task-specific branches at higher scales.

Decoders: Homography Outputs

At each scale, the homography estimation modules calculate four offsets, di-

rectly linked to homographies through the Direct Linear Transformation, as demonstrated in previous works[78, 81]. The motion-gated attention modules then engage by multiplying the features g_t^s with segmentations M_t^s to selectively filter out context irrelevant to the target motion. Thereafter, the channel dimensionality undergoes halving through a 3×3 convolution, while a second one focuses on extracting features from the restored output R_t^s. The combination of these two streams results in the formation of features h_t^s.

At each scale, these features h_t^s and $W_{\tilde{H}^s}(h_{t-1}^s)$ are utilized to predict offsets through shallow downstream networks. The predicted offsets at lower scales are then transformed back into homographies and cascaded bottom-up [81] to refine the higher-scale ones. Following a methodology similar to [78], we implement blocks of 3×3 convolutions, coupled with ReLU activations, batch normalization, and max-pooling to decrease the spatial size of the features. A dropout of 0.2 is applied just before the regression layer. For $s = 1$, the convolution output channels are set as 64, 128, 256, 256, and 256. For $s = 2$ and $s = 3$, the network depth is truncated from the second and third layers onward, respectively.

6.2.2 Loss Function

In this Subsection, we describe the two components that address the third and fourth hypothesis of this dissertation. Regarding the former, we assume that not all scales are necessary for all tasks. Even more, we hypothesize that when all tasks exist at all scales, gradient conflicts naturally occur and harm multi-task performance. We propose NSS to learn whether all gradients are really necessary or whether a few of them actually do not contribute significantly to the multi-task loss minimization. If gradients at some scale do not improve significantly performance of its task, then the task is omitted at this scale, thus minimizing gradient conflicts and freeing up parameters. Regarding the latter, we attempt to mitigate the diverse training paces issue by fixing the task-wise training paces to be equal by simply exploiting the numerical loss values, that is, without costly access to the per-task gradient vectors.

Neural Scale Search

We revisit the MOST output space of Figure 6.3. Each task is estimated at multiple scales via the multi-scale property of our network. Each scale is assumed to yield some performance improvement and conversely, lower scale outputs yield some estimation error compared to the higher ones. Let us denote the estimation error of the output upsampled from the lowest scale to the highest as follows:

$$\epsilon_i^S = \mathcal{L}_i \left(\mathcal{O}_i^1, u_i^{S,1}(\hat{\mathcal{O}}_i^S) \right), \tag{6.1}$$

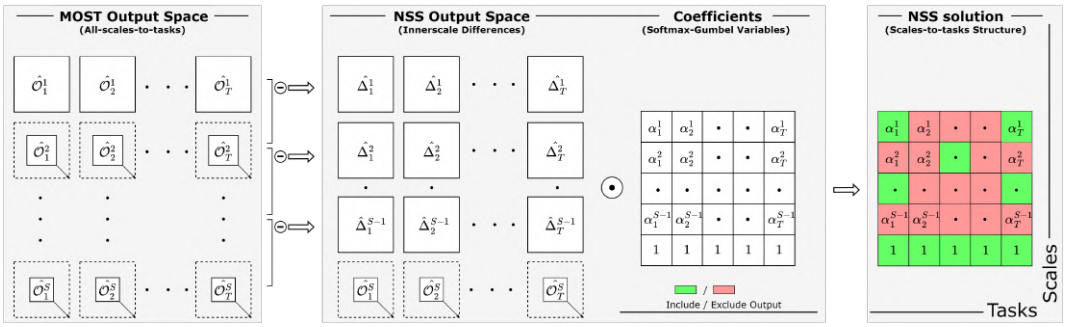


Figure 6.3: The output space of a MOST network entails multiple task outputs at multiple scales. NSS optimizes the task scaling by coupling the innerscale differences with learnable coefficients in the learning objective, to derive a more efficient architecture, illustrated as the NSS solution.

In NSS we assume that, for a well-defined architecture, the output at higher scale $s-1$ should improve or, at least, not degrade the estimation of $\mathcal{O}_i$ compared to a lower scale s. Subsequently, for $s > S$ and for all tasks i, the following holds:

$$\epsilon_i^{s-1} \leq \epsilon_i^s \tag{6.2}$$

where:

$$\epsilon_i^s(\omega, \theta) = \mathcal{L}_i \left(\mathcal{O}_i^1, u_i^{S,1}(\hat{\mathcal{O}}_i^S(\theta)) + \frac{1}{S-1} \sum_{j=S-1}^{s} \omega_i^s \Delta_i^s(\theta) \right),$$

and ω_i^s is some weighting scalar and

$$\Delta_i^s = u_i^{s,1} \left(u_i^{s,s+1}(\hat{\mathcal{O}}_i^{s+1}) - \hat{\mathcal{O}}_i^s \right) \tag{6.3}$$

As depicted in Figure 6.3, we evaluate the innerscale output differences Δ_i^s to gauge the impact of scale s on the i-th task. In essence, each difference signifies the influence of each additional scale compared to its preceding scale, with the lowest scale serving as the baseline and remaining constant. Let $\alpha \in (0,1)^{S \times T}$ be a matrix of probabilities with coefficient $\{\alpha_i^s\}_{i,s=1}^{T,S}$ quantifying whether performance improvement brought by Δ_i^s across tasks $\mathcal{O}_i^s$ is significant, such that:

$$\begin{aligned} \mathcal{O}_i^s &\in \mathcal{F}_{\theta,\alpha} \quad \text{if} \quad \alpha_i^s \to 1 \\ \mathcal{O}_i^s &\notin \mathcal{F}_{\theta,\alpha} \quad \text{if} \quad \alpha_i^s \to 0 \end{aligned} \tag{6.4}$$

In this context, the optimization of the network structure $\mathcal{F}_{\theta,\alpha}$ can be achieved through the following optimization problem. This problem involves the simultaneous learning of the optimal scales-to-task structure α for a network $\mathcal{F}$ and the network weights θ:

$$\mathcal{L}_{NSS}(\theta, \alpha) = \sum_{i}^{T} \lambda_i \epsilon_i^1(\alpha, \theta) + \mathcal{L}_0(\alpha), \tag{6.5}$$

Here, $\mathcal{L}_0$ represents a sparsity measure [30], such as $\mathcal{L}_0(\alpha) = \sum_s \sum_i \log(\alpha_i^s)$. The optimization problem in Equation (6.5) can be effectively addressed using the Softmax-Gumbel scheme[31]. This scheme enables the direct and single-level optimization of the decision structure concurrently with the network parameters.

Adaptive Task Balancing

Training a multi-task network typically involves minimizing the sum of a linear combination of task-specific losses (linear scalarization), as expressed in Equation (5.3) or Equation (6.5). However, our fourth hypothesis suggests that the task-specific loss weights, which initially balance the tasks during early training stages, may lose this balance in later stages. Previous research relies on access to internal network gradients [32] or hyperparameter optimization [19]. In our approach, assuming a training process with a total of K weight updates, we propose adaptively adjusting the task-specific loss weights ${\lambda_i}^k$ at each weight update $k = 1, \ldots, K$ after the forward pass, where access to the numerical loss values is available. Specifically, we impose constraints on the task-specific numerical loss values v_i^k, ensuring they are equal in pairs, and the sum of their values retains its magnitude, V^k, after reweighting. In other words, for all $i \in T$ and $k \in K$, the updated weights λ_i^k are determined as the solution to the following linear system:

$$\begin{cases} \sum_{i}^{T} \lambda_i v_i^k = V^k \\ \forall i, j \in T \qquad \lambda_i v_i^k = \lambda_j v_j^k \end{cases} \tag{6.6}$$

where $v_i^k, V^k \in \mathcal{R}^+$, $v_i^k = \mathcal{L}_i\left(\mathcal{O}_i^s, \hat{\mathcal{O}}_i^s(\theta^{k-1})\right)$ and $V^k = \sum_i^T v_i^k$. Given the linearity of the system, it can be reformulated in matrix form and easily solved using Gaussian elimination. This solution ensures that tasks are trained at a consistent speed, maintaining a rate of change for task-specific losses of 1 across time steps. Importantly, it preserves the magnitude of the multi-task loss V^k, with no influence on the multi-task learning rate and no need for initialization.

6.3 Experiments

6.3.1 Dataset

In this Section, we initially discuss the dataset, consisting of 100 intra-oral sequences from real dental interventions and their partition into training, validation, and test sets, consisting of 65, 10, and 25 videos, respectively. The distribution of frames in each set is 49K for training, 4K for validation, and 17K for the test set. The dataset specifics, including the count of manual frame segmentations and per-task baseline metrics, are summarized in Table 6.1. The general pipeline for data

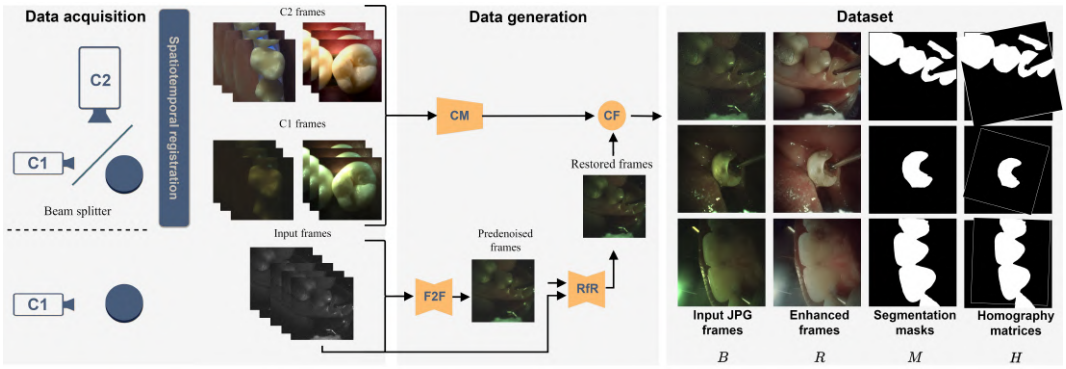


Figure 6.4: Data acquisition and generation pipeline for the publicly available Vident-Real-100 dataset. Top branch:We use a beam splitter in phantom (PH) scenes to acquire pairs of video frames and learn a color mapping network (CM).The learnt color function (CF) is applied on the restored frames, outputted from the bottom branch.Bottom branch: We acquire video sequences in real intra-oralscenes (R). The frames are processed for noise removal and sharpening, before being passed onto the CF component to colorize them. (b) Examples of three videosnippets from the dataset

acquisition and label generation is depicted in Figure 6.4.

Data	Train	Validation	Test
Videos	65	10	25
Frames	49K	4K	17K
Segmentations	426	40	138
PSNR	18.03	18.84	17.39
SSIM	0.834	0.857	0.821
MACE	11.48	7.59	9.29
IoU	0.208	0.206	0.254

Table 6.1: Summary of the Vident-real-100 dataset.

Noise, blur, colorization

For the enhancement of dental video data in terms of noise, blur, and colorization, we employ a dual-branch methodology. In the initial branch, we integrate a dental microcamera (C_1) with a larger, high-quality camera (C_2), both firmly coupled via a 50-50 beam splitter to simultaneously capture phantom scenes featuring natural teeth (PH). A color mapping (CM) network is trained to approximate the colors of C_2 via C_1, utilizing the spatiotemporally aligned frames from both cameras. The learned color mapping function (CF) is then applied to enhance the color characteristics of frames captured solely by C_1. The second branch, exclusive to camera C_1, involves unsupervised denoising in the RAW domain using the *Frame2Frame* (F2F) approach [161]. Subsequently, a "restore-from-restored" (RfromR) method [168] is employed, wherein the pre-denoised frames are subjected to additional defocus blur. The model of Lee et al. [168] is trained to remove noise, blocking artifacts, and enhance the sharpness of the input data. The restored frames (R) are obtained by

applying the color mapping function from the first branch to the output sequences of the second branch.

Segmentation masks and homographies

In the context of segmentation masks and homographies, denoted as M and H respectively, we manually annotate teeth on select frames from each video. A pre-trained HRNet48 [163] is then fine-tuned on these annotated frames to facilitate automatic teeth segmentation across the entire dataset. Optical flows (OF) between consecutive frames are computed using RAFT [84], fine-tuned on synthetic dental data to align with the motion characteristics of intra-oral sequences. The motion fields are subsequently cropped with teeth masks (M) to eliminate irrelevant object motion. Finally, a partial affine homography (H) is fitted using RANSAC to the segmented motion field; the derived H homography consists our ground truth label for the homography estimation task.

6.3.2 Setup

We perform multiple experiments to verify the validity of the novel *MOST-NET++* architectural instantiation. Thereafter, we perform diagnostics and an extensive performance evaluation for NSS (third hypothesis) and ATB (fourth hypothesis).

- We reconfigure *MOST-NET++* by removing different branches to showcase that the proposed design is maximally synergic and thus the most optimal.
- We assess the NSS training routines, the performance of the discovered *MOST-NSS++* architecture, the gradient conflicts, and the scale-wise performance improvements compared to *MOST-NET++*.
- We compare ATB with efficient baselines for different MOST architectures and discuss the training curves and the generalization performance.
- We compare *MOST-NET++* and *MOST-NSS++* with single task state-of-the-art methods for restoration (ESTRNN [48], MIMO-UNet [49]), homography estimation (MHN [81]) and binary segmentation (DLV3+ [90], UNET++ [89]) and multi-task architectures ([165]).

We train MOST-NET++ with different loss functions: the Charbonnier loss [158] as $\mathcal{L}_1$, binary cross-entropy [159] as $\mathcal{L}_2$, and Mean Average Corner Error (MACE) [78] as $\mathcal{L}_3$. Task-specific loss weights, denoted as linear scalarization (LS), are manually set to 1×10^{-1}, 2×10^{-1}, and 1.4×10^{3} for λ_1, λ_2, and λ_3 respectively. Regarding MOST-NSS++, the initial phase involves pretraining the network weights θ over the training dataset $\mathcal{D}$ for 40 epochs. Subsequently, the search phase starts, optimizing both the scale architecture α and the network weights θ jointly through a single-level optimization scheme, accomplished in a single backward pass. The

search phase spans two epochs. Upon derivation of the optimal architecture, we eliminate task-specific decoders for scales where tasks are not performed and retrain the network from scratch, as per the common Neural Architecture Search (NAS) practices.

Throughout all experiments, a batch size of 5 is utilized, and Adam serves as the optimizer with a learning rate of 1×10^{-4} for θ, reduced to 1×10^{-6} with cosine annealing. For α, a learning rate of 1×10^{-2} is employed to encourage more exploration. We use horizontal and vertical flips with a probability of 0.5, random channel perturbations, and color jittering, for data augmentation. Consistent configurations are maintained across all models to ensure a fair comparison, including ablations, comparisons, and diagnostics. The experiments are conducted using PyTorch 1.10 on 2×NVidia V100, and the frames per second (FPS) metric is measured on NVidia RTX 5000 to approximate the deployment environment. Performance evaluation encompasses video restoration, involving frame-wise PSNR and SSIM, homography estimation with Mean Average Corner Error (MACE), and semantic segmentation utilizing Intersection over Union (IoU) for teeth.

6.3.3 Results

Synergy

To gauge the impact of auxiliary tasks on the overall *MOST-NET++* architecture, we begin by evaluating architectural synergy. Firstly, we remove alignment by excluding the homography warping (W) operation from the feature alignment module, warping it solely with the identity matrix. Secondly, we eliminate the segmentation mask from the homography decoders (S) to examine the significance of enabling the teeth segmentation map to interact with other tasks. While both experiments yield consistent performance in segmentation and homography estimation, Table 6.2 highlights a notable degradation in image quality, resulting in a decrease of 1.7 to even about 2.2 dB. This underscores the importance of aligning frames and emphasizing the teeth region for optimal restoration results.

Architecture	S	W	PSNR ↑	SSIM ↑	MACE ↓	IoU ↑	#P(M) ↓	FPS ↑
MOST-NET++	✓	✗	31.47 (32.12)	0.976 (0.972)	5.78 (4.81)	0.627 (0.609)	8.9	5.0
MOST-NET++	✗	✓	32.01 (31.68)	0.979 (0.973)	**5.56** (5.36)	0.631 (0.548)	8.9	5.0
MOST-NET++	✓	✓	**33.69** (32.92)	**0.984** (0.980)	5.85 (5.14)	**0.643** (0.597)	8.9	5.0

Table 6.2: Ablation study for using segmentation (S) and homography (W) outputs in *MOST-NET++* on the test (validation) set. Eliminating task interactions by removing either H_t or M_t results in significant, i.e. 1.5-2.2dB, performance degradation for video restoration, demonstrating the synergy of the proposed architecture.

Neural Scale Search

NSS reaches convergence, as depicted in Figure 6.7, where we present the convergence diagnostics for a temperature of 5, averaged across three runs. The highest scale is allocated for the restoration task, emphasizing the need for a finer level of detail. The medium scale is assigned to segmentation, considering that

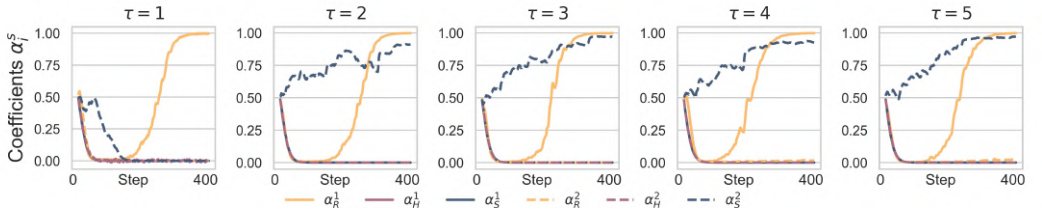


Figure 6.5: The same scales-to-tasks structure is derived for all temperatures $\tau > 1$ validating the robustness of the NSS solution. When $\tau = 1$, the solution is even sparser, retaining the highest scale only for the restoration task, while performing segmentation and homography estimation solely at the lowest scale.

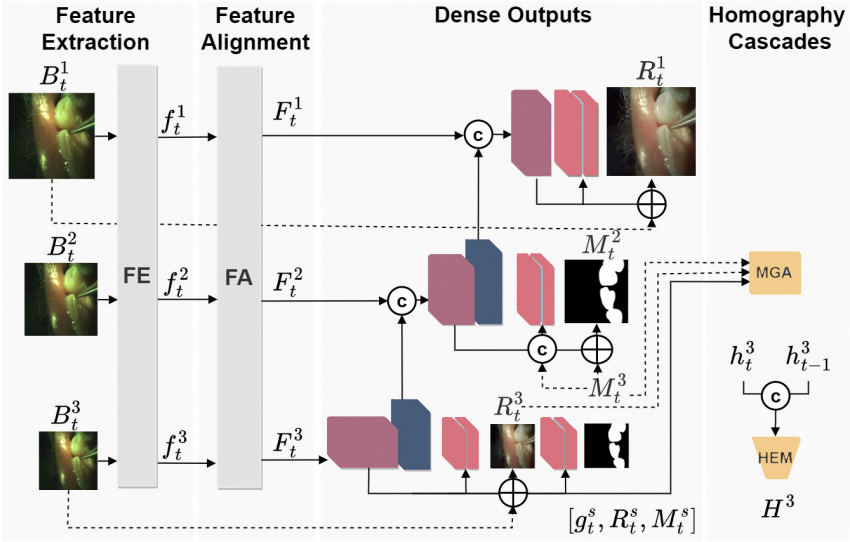


Figure 6.6: The reduction of the *MOST-NET++* architecture into *MOST-NSS++*. Similarly to Figure 6.2, feature extraction and alignment remain identical. However, by omitting task-specific decoders at multiple scales, *MOST-NSS++* is more lightweight. *MOST-NSS++* performs all tasks at the lowest scale as shown by the outputs (R^3, M^3, H^3) while it retains the middle scale for segmentation (M^2) and the highest scale for restoration (R^1).

predictions are already robust, and performance improvement on the high scale is negligible due to the sparsity constraint. This scales-to-tasks structure aligns with the design principles observed in modern single-task networks. The restructured, discovered architecture is depicted in Figure 6.6.

It's noteworthy that NSS achieves convergence to the same solution in all runs, without employing temperature annealing as seen in prior works such as [30, 21]. Additionally, the resulting configuration is derived for various temperature parameters, as illustrated in Figure 6.5.

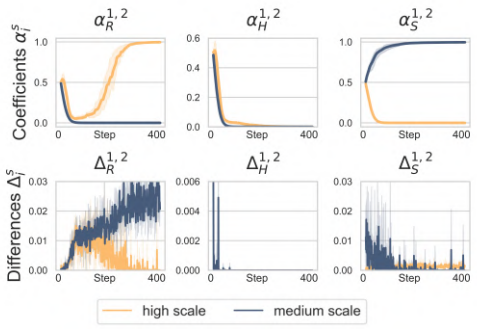


Figure 6.7: Convergence diagnostics of the α_i^s and Δ_i^s variables for a temperature of $\tau = 5$, across three NSS runs.

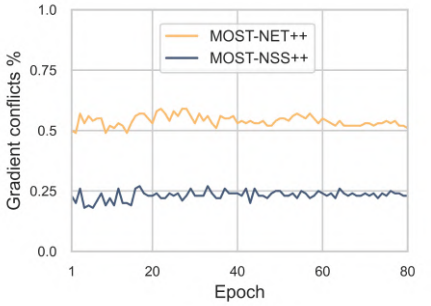


Figure 6.8: Percentage of multi-task gradient conflicts with respect to the total number of updates per epoch for *MOST-NET++* and *MOST-NSS++* on a batch size of 5.

Upon reaching convergence, we conduct a retraining of the discovered *MOST-NSS++* network from scratch. Interestingly, the resulting architecture demonstrates more efficient training dynamics, as depicted in Figure 6.9, in contrast to *MOST-NET++* or*MOST-NET*, which tend to plateau faster irrspectively of the optimization scheme employed.

Subsequently, we conduct a comparative analysis of *MOST-NET++* and *MOST-NSS++* in terms of conflicts of gradients [35] across epochs. These conflicts arise when a pair of gradient vectors, computed from the losses of two different tasks, produces an angle higher than 90 degrees. In Figure 6.8, we illustrate that while conflicts in task-specific gradients occur in approximately over 55% of weight updates in MOST, this percentage dramatically decreases to 25% for *MOST-NSS++*. This suggests that the performance improvement of NSS can be attributed not only to the discovered optimal scales-to-tasks structure but also to the introduced sparsity, effectively reducing negative transfer.

Lastly, in Figure 6.10, we showcase the performance improvement across scales on *MOST-NET++* compared to *MOST-NSS++* and observe that the latter utilizes its parameter resources more efficiently. Particularly, *MOST-NSS++* outperforms *MOST-NET++* on all tasks at the lowest scale, highlighting the effectiveness of NSS in resource allocation. Shared scales (lowest) exploit the gradient conflict mitigation to train faster and reach a better optimum while task-specific scales (medium, high) specialize for each task and, overall, train a more effective architecture.

Adaptive Task Balancing

We conduct a comparative analysis of ATB against methods that do not require access to internal network gradients. We employ the baseline alternatives DWA, RWL, and LS. RWL attempts to bypass bad local optima issues by passing the uniformly-sampled loss weights from a softmax function. Likewise it introduces temporally local asymmetries in task weighting which are however averaged over

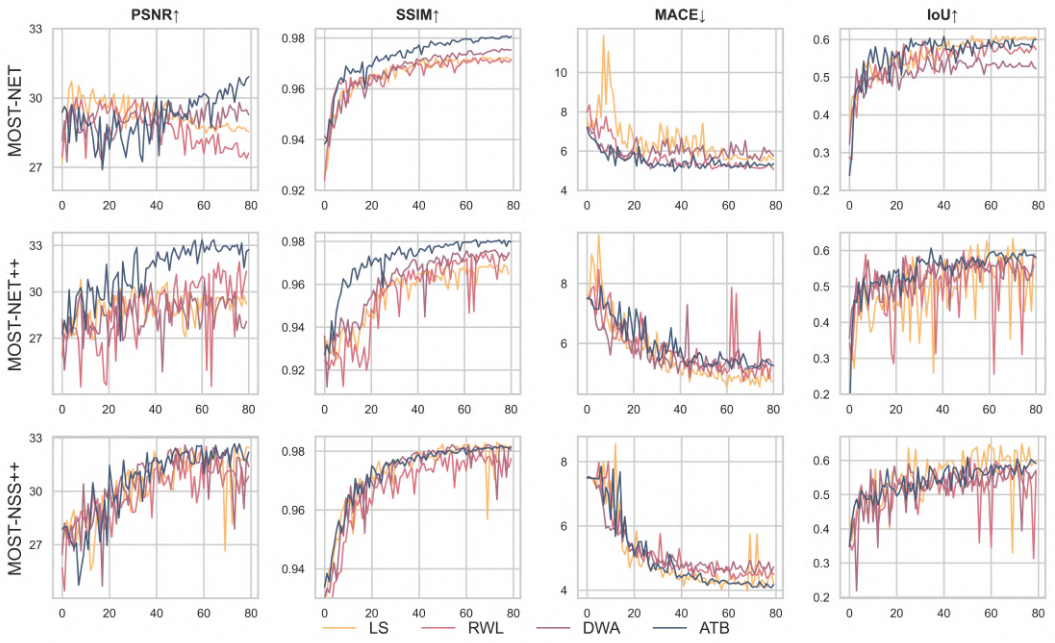


Figure 6.9: ATB compared to DWA, RWL and LS on the validation set for three different MOST architectures. ATB yields smoother curves on the validation set indicating that optimization states revolve around flatter regions of the feature space.

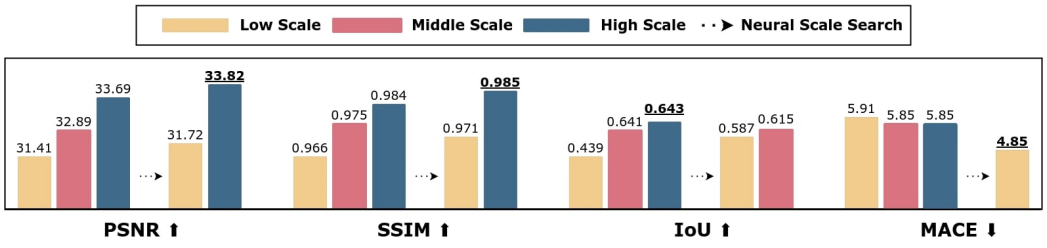


Figure 6.10: Performance across scales for *MOST-NET++* (all-scales-to-all-tasks) and *MOST-NSS++* (discovered scales-to-tasks) architectures. Each metric is illustrated with three bars for *MOST-NET++* (left side of the arrows), which predicts one output per task and scale. *MOST-NSS++* (right side of the arrows) on the contrary bears less bars as not all tasks at all scales. *MOST-NSS++* improves performance on most metrics despite the lower parameter count and faster runtimes.

time since the sampling distribution is uniform. RWL incurs zero cost comparatively and has shown to be a competitive alternative to multi-task optimizations methods. DWA addresses the pet-task training rates explicitly, by defining new per-task loss weights based on the numeric loss values and the per-task, relative descending rate. LS refers to simple linear scalarization, which attempts to bridge the magnitudes of the per-task gradients. We repeat those experiments across three MOST architectures: MOST-NET [165], the proposed MOST-NET++, and the optimized *MOST-NSS++*.

Across various runs and tasks, ATB consistently outperforms the other methods, as detailed in Table 6.3. Notably, ATB demonstrates better performance and contributes to more stable training, as evidenced by the smoothness of the training curves. It is worth noting that, among the MOST architectures, *MOST-NSS++* exhibits the least sensitivity to different training regimes, highlighting the effectiveness of its design.

Our experiments reveal that while both DWA and RWL generally outperform LS, they can exhibit instability. It is further evident that, DWA shows better training curves than RWL. DWA addresses the pet-task training rates explicitly. RWL on the other hand lacks explicit handling of diverse training speeds. In our experiments we find it to be less stable than simple linear scalarization.

Architecture	**Method**	**PSNR ↑**	**SSIM ↑**	**MACE ↓**	**IoU ↑**	**#P(M) ↓**	**FPS ↑**
MOST-NET	LS	31.14 (28.72)	0.977 (0.972)	6.47 (5.48)	0.599 (0.604)	8.7	5.0
MOST-NET	RWL	30.48 (29.94)	0.973 (0.969)	6.80 (5.63)	0.624 (0.595)	8.7	5.0
MOST-NET	DWA	31.57 (29.71)	0.977 (0.971)	6.26 (5.57)	0.598 (0.542)	8.7	5.0
MOST-NET	ATB	**32.55** (30.82)	**0.984** (0.980)	**5.94** (5.15)	**0.649** (0.616)	8.7	5.0
MOST-NET++	LS	31.24 (29.07)	0.976 (0.969)	5.51 (4.51)	**0.644** (0.604)	8.9	5.0
MOST-NET++	RWL	33.21 (31.99)	0.979 (0.974)	5.62 (4.73)	0.635 (0.586)	8.9	5.0
MOST-NET++	DWA	32.81 (30.71)	0.979 (0.976)	**5.44** (5.53)	0.642 (0.588)	8.9	5.0
MOST-NET++	ATB	**33.69** (32.92)	**0.984** (0.980)	5.85 (5.14)	0.643 (0.597)	8.9	5.0
MOST-NSS++	LS	33.56 (32.45)	0.984 (0.982)	5.40 (4.32)	**0.641** (0.612)	6.8	6.4
MOST-NSS++	RWL	33.07 (32.51)	0.983(0.981)	5.21 (4.34)	0.628 (0.585)	6.8	6.4
MOST-NSS++	DWA	33.78 (32.00)	**0.985** (0.982)	5.07 (4.50)	0.626 (0.585)	6.8	6.4
MOST-NSS++	ATB	**33.82** (32.66)	**0.985** (0.982)	**4.85** (3.85)	0.615 (0.608)	6.8	6.4

Table 6.3: Performance evaluation for ATB against DWA, RWL and LS on the test (validation) set for three different MOST architectures. ATB outperforms compared methods for most tasks and architectures, and generalizes better on the test set.

Comparisons

We further showcase the performance of the final solutions on the dental application by comparing them to established state-of-the-art single-task models, as summarized in Table 6.4. We adapt *MIMO-UNet* and *ESTRNN*, specifically designed for restoration tasks, and integrate them into our framework. Additionally, we incorporate *DeepLabv3+* and *UNET++* for semantic segmentation, and implement *HMG* for homography estimation, successfully reproducing the MS-COCO results reported by the authors. Our baseline metrics involve blindly predicting restored images, homography, and segmentation using input frames, an identity matrix (representing no motion), and assigning all pixels to teeth classes.

MOST-NSS++ demonstrates superior performance compared to the restora-

tion networks, with substantial improvements. They perform competitively, if not slightly lower, in homography estimation and semantic segmentation. *MOST-NSS++* outperforms *MOST-NET++* by a large margin on restoration and homography estimation while performing slightly worse only for segmentation, with less parameters and better runtimes. Notably, *MOST-NSS++* achieves further performance enhancement over *MOST-NET++* while employing 24% fewer parameters and achieving a 28% increase in frames per second (FPS).

Architecture	**PSNR** ↑	**SSIM** ↑	**MACE** ↓	**IoU** ↑	**#P(M)** ↓	**FPS** ↑
BASELINE	17.39 (18.84)	0.821 (0.857)	9.19 (7.59)	0.254 (0.206)	–	–
HMG [81]	–	–	4.24 (3.78)	–	6.2	28.4
MIMO-UNet [49]	29.18 (29.54)	0.977 (0.975)	–	–	6.8	4.6
ESTRNN [48]	33.08 (32.94)	0.980 (0.981)	–	–	2.3	10.6
UNET++ [89]	–	–	–	0.695 (0.762)	50.0	7.9
DL [90]	–	–	–	0.680 (0.745)	26.7	25.5
MOST-NET [165]	32.55 (30.82)	0.984 (0.980)	5.94 (5.15)	**0.649** (0.616)	8.7	5.0
MOST-NET++	33.69 (32.92)	0.984 (0.980)	5.85 (5.14)	0.643(0.597)	8.9	5.0
MOST-NSS++	**33.82** (32.66)	**0.985** (0.982)	**4.85** (3.85)	0.615 (0.608)	**6.8**	**6.4**

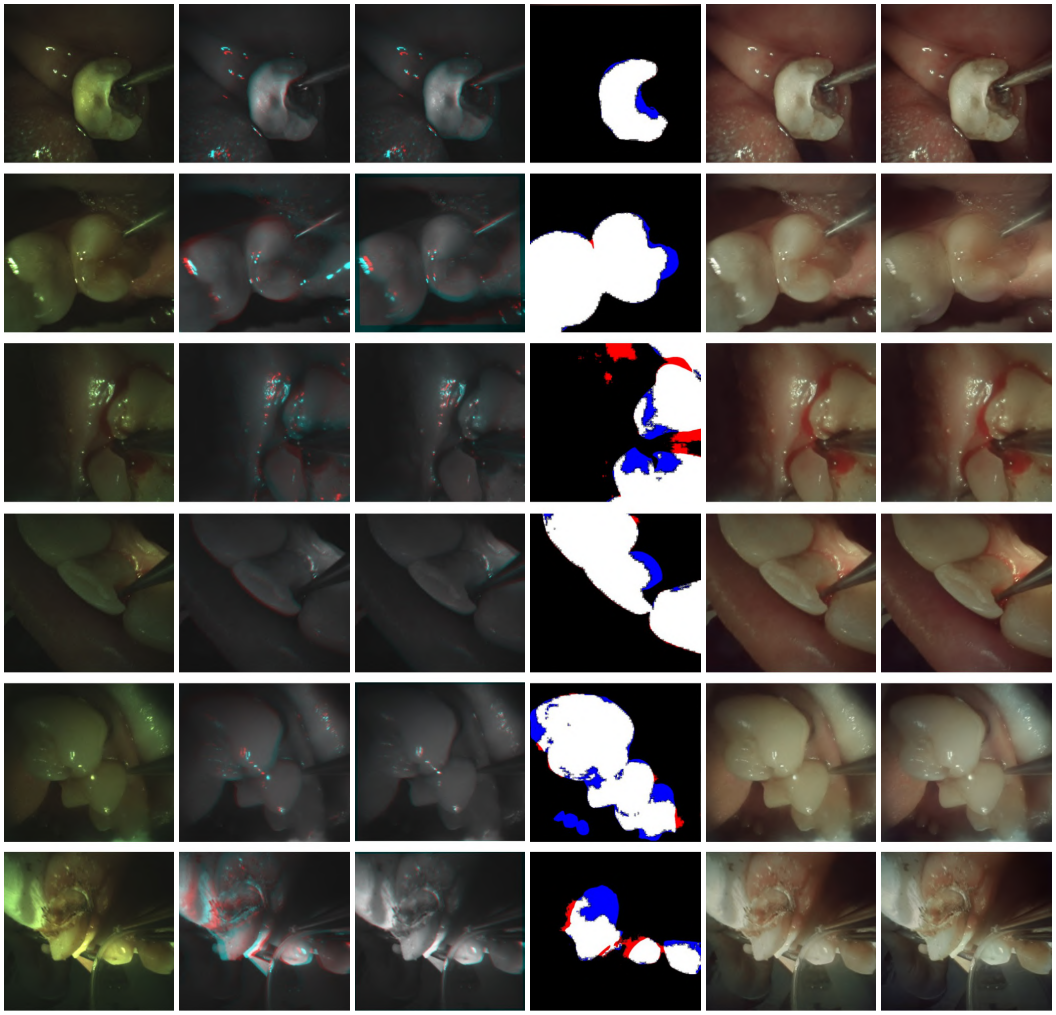
Table 6.4: Performance evaluation of *MOST-NET++* and *MOST-NSS++* against single- and multi-task networks on the test (validation) set. The proposed network and its reduced variant outperform previous multi-task work. They further perform competitively, or even better than state-of-the-art single-taskers, at lower computational resources.

Qualitative Results

We further illustrate the qualitative results of this experimental study in Figure 6.11. From left to right, columns depict the previous and current input frames, the previous frame overlaid on the current frame without alignment, i.e. warped with the identity matrix (large differences are illustrated with red and cyan colors), the homography-warped previous frame, overlaid on the current one (smaller red and cyan color intensities show better alignment), the output segmentation mask, and the restored current frame. The network yields vivid colors with less noise and blur compared to the input frames. Simultaneously, it can generally segment the masks (rows 1,2,4 and 6) but fails on very dark imagery (row 5) when artifacts such as blood (row 3) or excessive water appear. Last, the homography estimations are satisfying for small inter-frame motion but fail when the scale component becomes larger since the variance in the motion fields increases, and the optimal homography is compromised. Row 6 shows a successful example where the misalignment of column 2 is drastically recovered in column 3. Row 3 further shows good results where the network manages to align the frames regardless of the tooth deformation caused by the water.

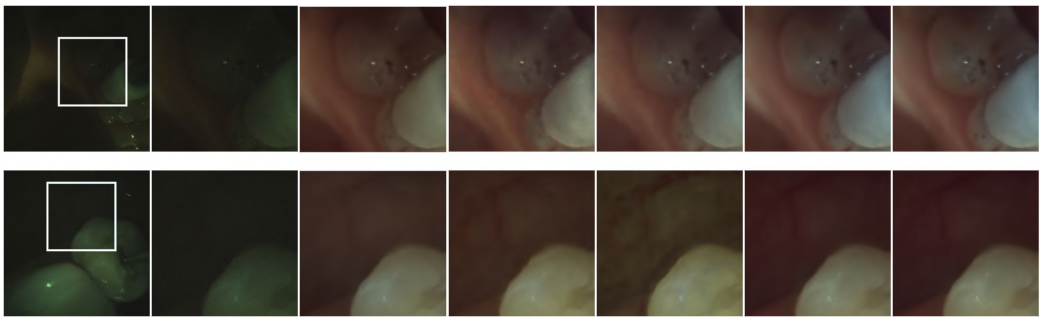
Thereafter, we visualize the qualitative results of video restoration compared to other state-of-the-art methods in Figure 6.12. Despite accommodating multiple tasks, *MOST-NSS++* outputs better colors and sharper edges compared to *MOST-NET* and popular, state-of-the-art single-task networks trained for the same task. The differences are clearly reflected on the PSNR metric.

Last, we showcase how *MOST-NSS++* improves performance with each ad-



(a) B_t (b) $W_I(B_{t-1})$ (c) $W_{\hat{H}}(B_{t-1})$ (d) $\hat{M}_t$ (e) $\hat{R}_t$ (f) R_t

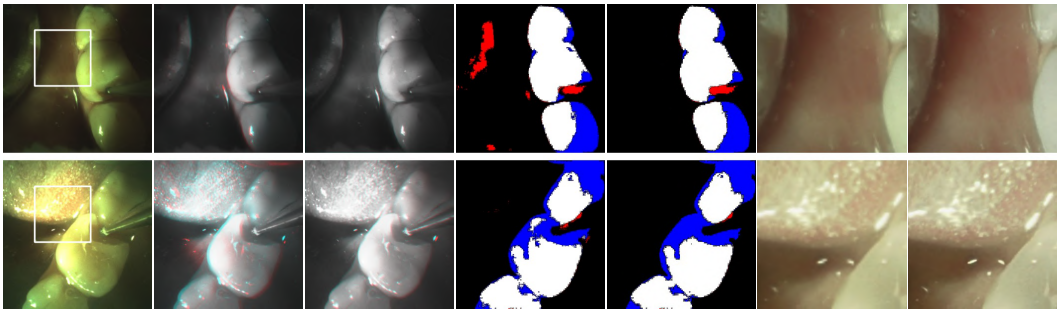
Figure 6.11: Qualitative results of *MOST-NSS++* on six real intra-oral video sequences. We showcase the a) current input frame, b) unaligned consecutive frames with the past frame overlaid on the current as $W_I(B_{t-1})$, c) their homography-aligned version as $W_{\hat{H}}(B_{t-1})$, d) segmentation map $\hat{M}_t$ where blue and red denote the FP and FN pixels respectively, e) restored frame $\hat{R}_t$ and f) its GT label R_t .



(a) Frame Top / Bottom PNSR	(b) Patch 15.1 dB / 19.1 dB	(c) MIMO-UNet 24.9 dB / 29.8 dB	(d) ESTRNN 33.2 dB / 31.3 dB	(e) MOST-NET 30.2 dB / 28.2 dB	(f) MOST-NSS++ **36.7 dB** / **35.6 dB**	(g) GT ∞/∞

Figure 6.12: Qualitative results for video restoration.

ditional scale of its derived architecture. The results are illustrated in Figure 6.13. Indeed, segmentation and restoration results substantially improve with each scale. Homography, on the contrary, bears no performance improvement and thus optimally uses only one scale, to free the next scales from noisy gradients and non-used parameters.



(a) B_t Top / Bottom / Metric	(b) W_I 11.21 / 9.72 / MACE	(c) $W_{\hat{H}^3}$ 3.31 / 1.94 / MACE	(d) M_t^3 0.678 / 0.621 / IoU	(e) M_t^2 0.731 / 0.632 / IoU	(f) R_t^3 30.16 / 31.97 / PSNR	(g) R_t^1 34.23 / 35.13 / PSNR

Figure 6.13: Qualitative performance improvement for *MOST-NSS++* across scales. Outputs at higher scales are more accurate, especially for video restoration where colors are more vivid (teeth are whiter) and frames are sharper (texture in the tongue and teeth contours are less blurry) at the highest scale R_t^1, compared to R_t^3. Segmentation maps are similarly refined at the middle scale M_t^2, compared to M_t^3, while homography estimation outputs achieve low errors already at the lowest scale eliminating the necessity for upscaling and thus reducing model parameters and minimizing gradient conflicts.

6.4 Discussion and Conclusion

In this Chapter, we employed a modified *MOST* architecture, *MOST-NET++*. The network was initially trained with the sum of the losses. Subsequently, it was trained with *NSS*, which discovered a new architecture, *MOST-NSS++*. The discovered network showed to improve in overall performance despite being essentially pruned. Specifically, *MOST-NSS++* achieves further performance enhancement over *MOST-NET++* while employing 24% fewer parameters and achieving a 28% increase in frames per second (FPS). We posit the improved network to the scales-to-tasks structure derived, which confirms our third hypothesis that not all scales are necessary for all tasks; even worse, such a setup introduces conflicts of gradients in about half of the model updates. We show that *MOST-NSS++* minimizes gradient conflicts and advances further from *MOST-NET++* to achieve very low parameter count, and even runtimes, despite accommodating a multitude of tasks with state-of-the-art performance.

Moreover we experimentally confirm that the *MOST* networks trained with *ATB* are consistently more stable in our experiments. We observe that they tend to plateau slower and generalize better on the test set. When most relevant methods are gradient-based, wherein the backward pass time scales linearly with the number of tasks, *ATB* bears almost no additional compute and is thus practical for real-world applications. In this Chapter, we experimentally confirm the fourth hypothesis, i.e. tasks train at uneven rates, and modelling those rates is beneficial for multi-task learning, for the application and studies at hand. The study in itself poses interesting questions for further investigation. How does *ATB* perform with scaling up the number of tasks? How does it adapt to more variant loss landscapes?

The application results demonstrate notable achievements across various aspects. The restoration process exhibits commendable performance, reaching the state-of-the-art for the dataset. Homography results display strong inter-frame accuracy, although a drift appears when aligning longer-range video frames. Segmentation, while showcasing relatively good performance compared to state-of-the-art, encounters more challenges, potentially stemming from the disparity in training, validation, or testing distributions. It is noteworthy that despite the dataset's abundant frames, each snippet displays and focuses on one or a few teeth, thus learning solid teeth representations is more challenging. To enhance representation learning, it is suggested that domain generalization or the incorporation of a vast array of data augmentations, learned end-to-end following the Neural Architecture Search paradigm, may be necessary. On the architectural design and at higher complexity, the integration of spatial attention mechanisms has proven to substantially elevate overall performance in architectural design.

In other notes with respect to this study, the trained *MOST-NET++* model was at an earlier stage of the work, reduced in 16bits of arithmetic precision via the TensorRT framework. The truncated model achieved runtimes of approximately 24 FPS, verifying its capability for real-time video processing applications. The work

described in this Chapter chapter builds upon the foundational research presented in a recent journal paper submission, incorporating and expanding upon its findings to provide a more comprehensive and nuanced exploration within the context of this doctoral thesis.

Chapter 7

Conclusions and Outlook

7.1 Exploring Multi-Task Architectures to Combine Visual Enhancement and Understanding

When multiple tasks are of interest, this thesis shows experimentally that they can be combined and jointly learnt in multi-task architectures. No matter the nature of the tasks, visual cues among tasks are shared to some extend. Even when the task outputs are less correlated, we hypothesized that there is some common knowledge overlap, which, if incorporated within a multi-task convolutional neural network appropriately, it can yield architectures with better performance-vs-compute trade-off compared to single task networks. Moreover, exploiting the dynamic nature of videos, provides abundant information that is typically disregarded in the multi-task literature.

In Chapter 3, we addressed two visual enhancement, that despite belonging to the same level of hierarchy, their outputs are less correlated. While the noise map is omnipresent on the whole image, deblurring in dynamic scenes with diverse depths is heavily space-variant. To study and tackle the tasks simultaneously we proposed R2-D4 to solve video denoising and deblurring simultaneously. We showed that we can achieve state-of-the-art performance on this challenging combination of visual scene enhancement tasks, even with very high levels of noise. The performance improvement over compared methods does originate from the architectural design as we show that even with lesser number of parameters and floating point operations, R2-D4 still performs better than end-to-end, single task networks. To a large extend, we posit the success of this architecture to the deformable offsets which function as accurate and robust motion predictors. The strong supervision from the multi-task signal is in itself enough to learn pixel displacements, align frames and effectively borrow context from neighbouring ones to restore even heavily degraded frame regions.

In Chapter 4, we addressed video deblurring and segmentation as representative tasks of visual scene enhancement and understanding respectively. Despite

belonging to a different level of hierarchy those two tasks are closely related. To address them, we proposed a multi-task learning architecture that leverages spatiotemporal features to handle motion. The proposed method achieves higher deblurring performance than its single task counterpart, and a reasonably high IoU score for dental instrument segmentation and runs at significantly faster runtimes than the combined single-task solutions. The multi-task network further achieves a $\times 4$ reduction in parameters compared to the system of the single-taskers, facilitating the deployment on smaller devices.

The experiments performed in Chapters 3 and 4 confirm that low- and mid-level visual scene enhancement and understanding tasks such as the likes of video deblurring, denoising and segmentation can be efficiently combined. Effectively, we propose two novel multi-task architectures, capable of accommodating tasks at faster runtimes or lower parameter count and GFLOPs compared to single task architectures. Therefore, we experimentally verify our first hypothesis.

Our experiments pave the way for further research. Initially, we confined our experimental focus to convolutional neural networks for a specific set of tasks. How might one introduce an additional, yet challenging visual scene enhancement task, such as video super-resolution, alongside denoising and deblurring? What would the findings of a similar study on transformers? While numerous visual enhancement tasks like shadow or reflection removal are increasingly relevant, their joint integration in efficient multi-task architectures remains unexplored in the existing literature. Another limitation of this thesis lies in its exclusive concentration on the RGB domain. The exploration of multitasking across various modalities related to medical imaging has been relatively overlooked. For instance, in the case of MRI (Magnetic Resonance Imaging) slices, Computed Tomography (CT), or X-ray scans. Integrating visual enhancement, such as signal denoising or contrast enhancement, with the comprehension of anomalies, such as tumor recognition, fracture identification, or other anatomical landmarks in medical images, could be learned concurrently. Just as we implicitly aligned frames via deformable offsets in Chapter 3, one could register CT, X-ray, or MRI images to facilitate precise localization and tracking of changes over time. Such experiments could extend the findings of this thesis to different domains and modalities.

7.2 Leveraging Multi-task Interactions Across Convolutional Scales

In Chapter 5, we introduced *MOST-NET*, a novel deep neural network designed for video processing, addressing various tasks across multiple scales. The architecture includes tasks ranging from low-level video restoration (deblurring, denoising, and color mapping) to mid-level tasks like homography-based motion estimation and teeth segmentation for dental scene understanding. The practical applicability of *MOST-NET* in computer-aided dental interventions is highlighted,

confirming the accommodation of diverse tasks in a multi-task architecture for RGB video scenes. Notably, *MOST-NET* achieves lower parameter count than combined, state-of-the-art, single-task models with fast runtimes, despite producing multiple task outputs per scale. The Chapter further emphasizes the importance of frame alignment, demonstrating how homography estimation as an auxiliary task can enhance the performance of other video tasks. To support further research in visual scene enhancement and understanding tasks, the *Vident-lab* video dataset, featuring natural teeth within phantom scenes, has been openly shared as a valuable resource in a domain with limited publicly available materials. The dataset is detailed in Katsaros et al. [165].

However, it is important to note some limitations. The *MOST-NET* architecture is confined, again, on convolutional neural networks in the RGB domain, and its experimental aspects primarily apply to macro-visualization environments. While exploring other domains would be intriguing, our current findings are constrained by the absence of relevant datasets. Additionally, despite integrating video denoising, deblurring, and color mapping within *MOST-NET*, all tasks are consolidated and modeled as a single visual scene enhancement task, following the findings of Chapter 3. Although this approach is effective, it raises questions about how the architecture scales when multiple outputs are provided. Would the decoupled gradient from the separated tasks be advantageous or detrimental to the other tasks? Lastly, exploring how the current multi-task convolutional architecture performs when adapted to a Transformer architecture is another interesting avenue. While attention blocks have proven valuable for cross-task information exchange, pre-training may be essential to match the performance of more mature and easily trainable convolutional architectures.

7.3 Searching for Multi-task Interactions Across Convolutional Scales

In Chapter 6, we proposed a new variant of *MOST-NET*, dubbed *MOST-NET++*, to improve the architectural components and the multi-task performance. However, we observed that performing all tasks at all scales is not always the best approach, as performance improvement across tasks is not uniform. Moreover, multi-task gradients generate a high number of conflicts and can lead to suboptimal solutions. To alleviate the issue, we searched for the optimal task interactions with *NSS*, which discovered a new architecture, *MOST-NSS++*. *NSS* relies on simple backpropagation and its gradients to determine the optimal structure instead of computationally expensive reinforcement learning or evolutionary approaches. Contrasted to other methods that require bi-level optimization, *NSS* converges stably across different runs and temperature hyper-parameters with single-level optimization.

The discovered network showed to improve in overall performance despite be-

ing essentially pruned. Specifically, *MOST-NSS++* achieves further performance enhancement over *MOST-NET++* while employing 24% fewer parameters and achieving a 28% increase in frames per second (FPS). We posit the improved network to the scales-to-tasks structure derived, which confirms our third hypothesis that not all scales are necessary for all tasks; even worse, such a setup introduces conflicts of gradients in about half of the model updates. We show that *MOST-NSS++* minimizes gradient conflicts and advances further from *MOST-NET++* to achieve very low parameter count, and even runtimes, despite accommodating a multitude of tasks with state-of-the-art performance. Those aforementioned experiments, verify the third hypothesis of this doctoral dissertation.

While *NSS* utilizes SNAS and the Softmax-Gumbel trick to acquire the optimal scales-to-tasks structure, more recent NAS methods demonstrate enhanced performance. DrNAS has garnered significant attention as an intriguing alternative. A thorough comparison between the current NSS approach based on DARTS, SNAS, and DrNAS, for example, could offer valuable insights into the learning dynamics of the MOST architecture and potentially enhance the acquired structure. Another compelling avenue for research involves conducting similar experiments on a Transformer-based MOST architecture to observe how these NAS methods perform with different architectures, benefiting from an improved gradient flow due to the increased number of neural connections introduced by the inherent attention mechanisms at each block.

7.4 Diverse Multi-task Training Speeds and Adaptive Task Balancing

This thesis engages with the issue of diverse multi-task training speeds. Progress across tasks often exhibits unevenness, leading some tasks' progress to dominate others, resulting in solutions that deviate significantly from the Pareto front. Existing approaches typically rely on computationally expensive access to internal network gradients, which scales linearly with the number of tasks. To mitigate this challenge, we proposed an alternative solution called *Adaptive Task Balancing* (ATB), where the loss weights are adjusted dynamically with each weight update.

The experimental results of Chapter 6 reaffirm that *MOST* networks trained using *ATB* consistently exhibit more stable training curves. They tend to plateau slower and showcase improved generalization on the test set. Importantly, *ATB* introduces almost negligible additional computational overhead compared to relevant methods, rendering it practical for real-world applications. The fourth hypothesis of this thesis that tasks train at uneven rates is empirically validated, and incorporating this understanding through *ATB* proves beneficial for multi-task learning in the application under consideration. This study prompts intriguing questions for further exploration, such as assessing the performance of *ATB* when scaling up the number of tasks and its adaptability to more diverse loss landscapes.

As stated above, *ATB* does not access the network gradients, to support training with large datasets and number of tasks. Other related methods focus solely on multi-task performance actually consider and access the multi-task network gradients, despite the memory requirements. However, the solutions typically treat the per-task network gradients as an entity, and try to stabilize training as such, i.e. on the whole network level. Another interesting line of research would be to attempt to stabilize multi-task learning by considering the per-task gradients on each individual layer or scale, thus offering more granularity on the modification of the multi-task training dynamics.

7.5 Contributions

To conclude, this thesis enumerates the contributions mentioned below. Initially, I experimentally confirmed that visual scene enhancement and understanding tasks can be effectively integrated within multi-task convolutional architectures. Specifically, video deblurring was learnt in pairs with video denoising and segmentation, with better performance than combined single task networks at lower compute. Therefore, two novel convolutional architectures were proposed. The first integrated the two challenging visual scene enhancement tasks together for the first time in the literature. The second addressed both visual enhancement and understanding tasks together, solving deblurring and segmentation concurrently. This confirmed the first hypothesis and was thoroughly discussed in Chapters 3 and 4.

Taking a step further, this thesis leveraged the multi-scale nature of convolutional architectures as expressed by its feature maps, to accommodate five different vision tasks of diverse hierarchies and scope. Specifically, a multi-task, multi-scale, multi-output architecture was introduced, allowing multiple tasks to interact and refine their predictions from the lowest scale, up to the highest. The assumptions, design and experimental findings were presented in Chapter 5 and verified the second hypothesis of this thesis.

Next, this thesis argued that embedding tasks within convolutional neural network parameters essentially necessitates different compute and granularity for each task. Therefore, I proposed Neural Scale Search to learn a more effective architecture; instead of manually designing the branch-out strategy of the multi-task decoders, NSS learnt them concurrently with the network weights. Chapter 6 discussed in part the NSS method, the experiments and the findings, and confirmed that the learnt structure was better as evidenced by higher performance and lesser multi-task gradient conflicts.

Last, this thesis introduced a simple yet effective training regime that experimentally outperformed related techniques and stabilized the multi-task training without costly access to internal network gradients. The relevant descriptions and experiments were discussed in parts of Chapter 6.

The contributions of this thesis were experimentally verified and part of the

experimental results supporting this thesis have been published in several scientific papers. In practical terms, the insights within this thesis can guide practitioners seeking efficient solutions for multiple tasks, proposing topological designs, architectural components, and multi-task training methodologies.

7.6 Future Work

The thesis primarily focuses on convolutional neural networks in the RGB domain, limiting exploration of other domains and modalities such as medical imaging datasets like MRI, CT, or X-ray scans. Integration of tasks across various modalities remains relatively unexplored. While this thesis successfully integrates tasks including but not limited to video denoising, deblurring, homography estimation and segmentation within multi-task architectures, questions arise regarding the scalability of the architecture when multiple outputs are provided. Naturally the gradient vectors grow in number with the number of the tasks, incurring gradient conflicts and compromising multi-task performance. Moreover, exploring the multi-task behaviour in Transformer architectures, and the necessity of pre-training for matching the performance of convolutional architectures, warrants further investigation. The thesis addresses the issue of uneven progress across tasks during multi-task training, proposing Adaptive Task Balancing (ATB) as a solution. However, questions remain about its adaptability to diverse loss landscapes, scalability with a larger number of tasks, and potential modifications to stabilize multi-task learning by considering per-task gradients on individual layers or scales for more granular modification of training dynamics.

Publications

This section lists all peer-reviewed journal and conference publications that the author has contributed during their Ph.D. studies. The published material referenced throughout this thesis are listed below:

- **Katsaros, E.**, Jezierska, A., Ostrowski, P. K., Lewandowska, E., Ruminski, J., & Wesierski, D. (2024). Multi-task Neural Scale Search for Intra-oral Interventions. To be submitted to the "Medical Image Analysis" journal [MNiSW **200 pts**]

- **Katsaros, E.**, Ostrowski, P. K., Włódarczak, K., Lewandowska, E., Ruminski, J., Siupka-Mróz, D., Lassmann, Ł., Jezierska, A., & Wesierski, D. (2022). Multi-task Video Enhancement for Dental Interventions. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 177-187). Springer, Cham. [MNiSW **140 pts**]

- **Katsaros, E.**, Ostrowski, P. K., Wesierski, D., & Jezierska, A. (2021). Concurrent video denoising and deblurring for dynamic scenes. IEEE Access, 9, 157437-157446. [MNiSW **100 pts**]

- **Katsaros, E.**, Jezierska, A., & Wesierski, D. (2021, July). Leveraging spatio-temporal features for joint deblurring and segmentation of instruments in dental video microscopy. In 2021 14th International Conference on Human System Interaction (HSI) (pp. 1-5). IEEE. [MNiSW **20 pts**]

- Ostrowski, P. K., **Katsaros, E.**, Wesierski, D., & Jezierska, A. (2022). BP-EVD: Forward Block-Output Propagation for Efficient Video Denoising. IEEE Transactions on Image Processing, 31, 3809-3824. [MNiSW **200 pts**]

- **Katsaros, E.**, Kopa Ostrowski, P., Jezierska, A., Lewandowska, E., Ruminski, J., & Wesierski, D. (2022). Vident-lab: a dataset for multi-task video processing of phantom dental scenes (1–) [dataset]. Gdańsk University of Technology. `https://doi.org/10.34808/1jby-ay90`

List of Figures

List of Tables

Bibliography

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA*, June 2016.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25, Lake Tahoe, USA*, pp. 1106–1114, Dec. 2012.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems, Montreal, Canada*, pp. 91–99, Dec. 2015.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA*, pp. 779–788, June 2016.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[9] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*,

"Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[10] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.*, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," *arXiv preprint arXiv:1712.01815*, 2017.

[11] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, *et al.*, "Dota 2 with large scale deep reinforcement learning," *arXiv preprint arXiv:1912.06680*, 2019.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[13] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, *et al.*, "Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic acids research*, vol. 50, no. D1, pp. D439–D444, 2022.

[14] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[15] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[17] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[19] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1871–1880, 2019.

[20] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *CVPR*, pp. 675–684, 2018.

[21] D. Brüggemann, M. Kanakis, A. Obukhov, S. Georgoulis, and L. Van Gool, "Exploring relational context for multi-task dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15869–15878, 2021.

[22] S. Vandenhende, S. Georgoulis, and L. V. Gool, "Mti-net: Multi-scale task interaction networks for multi-task learning," in *European Conference on Computer Vision*, pp. 527–543, Springer, 2020.

[23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[24] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1978, 2014.

[25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images.," *ECCV (5)*, vol. 7576, pp. 746–760, 2012.

[26] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6129–6138, 2017.

[27] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–308, 2009.

[28] G. Boracchi and A. Foi, "Modeling the performance of image restoration from motion blur," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3502–3517, 2012.

[29] E. Katsaros, P. Kopa Ostrowski, A. Jezierska, E. Lewandowska, J. Rumiński, and D. Wesierski, "Vident-lab: a dataset for multi-task video processing of phantom dental scenes," 2022.

[30] X. Sun, R. Panda, R. Feris, and K. Saenko, "Adashare: Learning what to share for efficient deep multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8728–8740, 2020.

[31] S. Xie, H. Zheng, C. Liu, and L. Lin, "Snas: stochastic neural architecture search," *arXiv preprint arXiv:1812.09926*, 2018.

[32] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *International conference on machine learning*, pp. 794–803, PMLR, 2018.

[33] L. Liu, Y. Li, Z. Kuang, J. Xue, Y. Chen, W. Yang, Q. Liao, and W. Zhang, "Towards impartial multi-task learning," in *International conference on learning representations*, iclr, 2021.

[34] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

[35] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.

[36] V. Kurin, A. De Palma, I. Kostrikov, S. Whiteson, and P. K. Mudigonda, "In defense of the unitary scalarization for deep multi-task learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12169–12183, 2022.

[37] A. Navon, A. Shamsian, I. Achituve, H. Maron, K. Kawaguchi, G. Chechik, and E. Fetaya, "Multi-task learning as a bargaining game," in *International Conference on Machine Learning*, pp. 16428–16446, PMLR, 2022.

[38] D. Xin, B. Ghorbani, J. Gilmer, A. Garg, and O. Firat, "Do current multi-task optimization methods in deep learning even help?," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13597–13609, 2022.

[39] B. Lin, Y. Feiyang, Y. Zhang, and I. Tsang, "Reasonable effectiveness of random weighting: A litmus test for multi-task learning," *Transactions on Machine Learning Research*, 2022.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[41] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 17–36, JMLR Workshop and Conference Proceedings, 2012.

[42] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *arXiv preprint arXiv:2006.05278*, 2020.

[43] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[46] M. Tassano, J. Delon, and T. Veit, "Fastdvdnet: Towards real-time deep video denoising without flow estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1354–1363, 2020.

[47] P. K. Ostrowski, E. Katsaros, D. Węsierski, and A. Jezierska, "Bp-evd: Forward block-output propagation for efficient video denoising," *IEEE Transactions on Image Processing*, vol. 31, pp. 3809–3824, 2022.

[48] Z. Zhong, Y. Gao, Y. Zheng, and B. Zheng, "Efficient spatio-temporal recurrent neural network for video deblurring," in *European Conference on Computer Vision*, pp. 191–207, 2020.

[49] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4641–4650, 2021.

[50] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren, "Spatio-temporal filter adaptive network for video deblurring," in *IEEE International Conference on Computer Vision*, pp. 2482–2491, 2019.

[51] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9396–9416, 2021.

[52] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17683–17693, 2022.

[53] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. Van Gool, "Vrt: A video restoration transformer," *arXiv preprint arXiv:2201.12288*, 2022.

[54] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3883–3891, 2017.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[56] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8174–8182, 2018.

[57] Y. Yuan, W. Su, and D. Ma, "Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3555–3564, 2020.

[58] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *IEEE International Conference on Computer Vision*, pp. 764–773, 2017.

[59] M. Suin, K. Purohit, and A. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3615, 2020.

[60] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14821–14831, 2021.

[61] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1279–1288, 2017.

[62] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *CVPR Workshops*, 2019.

[63] J. Pan, H. Bai, and J. Tang, "Cascaded deep video deblurring using temporal sharpness prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3043–3051, 2020.

[64] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[65] V. Santhanam, V. I. Morariu, and L. S. Davis, "Generalized deep image to image regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5609–5619, 2017.

[66] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[67] K. Zhang, W. Zuo, and L. Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4608–4622, 2018.

[68] X. Chen, L. Song, and X. Yang, "Deep rnns for video denoising," in *Applications of digital image processing XXXIX*, vol. 9971, pp. 573–582, SPIE, 2016.

[69] T. Vogels, F. Rousselle, B. McWilliams, G. Röthlin, A. Harvill, D. Adler, M. Meyer, and J. Novák, "Denoising with kernel prediction and asymmetric loss functions," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–15, 2018.

[70] A. Davy, T. Ehret, J.-M. Morel, P. Arias, and G. Facciolo, "Non-local video denoising by cnn," *arXiv preprint arXiv:1811.12758*, 2018.

[71] M. Tassano, J. Delon, and T. Veit, "Dvdnet: A fast network for deep video denoising," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1805–1809, IEEE, 2019.

[72] M. Claus and J. Van Gemert, "Videnn: Deep blind video denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

[73] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[74] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM international conference on multimedia*, pp. 1632–1640, 2019.

[75] F. Lv, Y. Li, and F. Lu, "Attention guided low-light image enhancement with a large scale low-light simulation dataset," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2175–2193, 2021.

[76] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1780–1789, 2020.

[77] L.-W. Wang, Z.-S. Liu, W.-C. Siu, and D. P. Lun, "Lightening network for low-light image enhancement," *IEEE Transactions on Image Processing*, vol. 29, pp. 7984–7996, 2020.

[78] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," *arXiv preprint arXiv:1606.03798*, 2016.

[79] C.-H. Chang, C.-N. Chou, and E. Y. Chang, "Clkn: Cascaded lucas-kanade networks for image alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2213–2221, 2017.

[80] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, pp. 674–679, 1981.

[81] H. Le, F. Liu, S. Zhang, and A. Agarwala, "Deep homography estimation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7652–7661, 2020.

[82] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, 2018.

[83] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, "Iterative deep homography estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1879–1888, 2022.

[84] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*, pp. 402–419, Springer, 2020.

[85] H. Gao, X. Tao, X. Shen, and J. Jia, "Dynamic scene deblurring with parameter selective sharing and nested skip connections," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3848–3856, 2019.

[86] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[87] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[88] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pp. 424–432, Springer, 2016.

[89] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11, Springer, 2018.

[90] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[91] S. Zhou, C. Li, and C. C. Loy, "Lednet: Joint low-light enhancement and deblurring in the dark," in *ECCV*, 2022.

[92] Y. Cui, C. Tang, and Q. Huang, "Joint face super-resolution and deblurring using multi-task feature fusion network," in *7th International Conference on Vision, Image and Signal Processing (ICVISP 2023)*, vol. 2023, pp. 57–61, 2023.

[93] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[94] M. Yu, M. Guo, S. Zhang, Y. Zhan, M. Zhao, T. Lukasiewicz, and Z. Xu, "Rirgan: An end-to-end lightweight multi-task learning method for brain mri super-resolution and denoising," *Computers in Biology and Medicine*, vol. 167, p. 107632, 2023.

[95] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Deep parametric 3d filters for joint video denoising and illumination enhancement in video super resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 3054–3062, 2023.

[96] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *CVPR*, pp. 4106–4115, 2019.

[97] H. Jung, Y. Kim, H. Jang, N. Ha, and K. Sohn, "Multi-task learning framework for motion estimation and dynamic scene deblurring," *IEEE Transactions on Image Processing*, vol. 30, pp. 8170–8183, 2021.

[98] J. Guo, H. Feng, H. Xu, W. Yu, and S. shuzhi Ge, "D3-net: integrated multi-task convolutional neural network for water surface deblurring, dehazing and object detection," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105558, 2023.

[99] S. Nazir, L. Vaquero, M. Mucientes, V. M. Brea, and D. Coltuc, "Depth estimation and image restoration by deep learning from defocused images," *IEEE Transactions on Computational Imaging*, 2023.

[100] T. Lin *et al.*, "Microsoft COCO: common objects in context," in *Computer Vision - ECCV - 13th European Conference, Zurich, Switzerland, Proceedings, Part V*, vol. 8693, pp. 740–755, Springer, Sept. 2014.

[101] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[102] K.-K. Maninis, I. Radosavovic, and I. Kokkinos, "Attentive single-tasking of multiple tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1851–1860, 2019.

[103] J. Vertens, A. Valada, and W. Burgard, "Smsnet: Semantic motion segmentation using deep convolutional neural networks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 582–589, IEEE, 2017.

[104] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[105] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3061–3070, 2015.

[106] M. Menze, C. Heipke, and A. Geiger, "Joint 3d estimation of vehicles and scene flow," *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, vol. 2, pp. 427–434, 2015.

[107] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, T. Darrell, *et al.*, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, vol. 2, no. 5, p. 6, 2018.

[108] S. Chennupati, G. Sistu, S. Yogamani, and S. A Rawashdeh, "Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[109] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 769–777, 2015.

[110] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. Van Den Hengel, and Q. Shi, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2319–2328, 2017.

[111] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Understanding deformable alignment in video super-resolution," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 973–981, 2021.

[112] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 261–270, 2017.

[113] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9308–9316, 2019.

[114] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 117–132, 2018.

[115] T. Isobe, S. Li, X. Jia, S. Yuan, G. Slabaugh, C. Xu, Y.-L. Li, S. Wang, and Q. Tian, "Video super-resolution with temporal group attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8008–8017, 2020.

[116] R. Feng, C. Li, H. Chen, S. Li, J. Gu, and C. C. Loy, "Generating aligned pseudo-supervision from non-aligned data for image restoration in under-display camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5013–5022, 2023.

[117] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3929–3938, 2017.

[118] J. Zhang, J. Pan, W.-S. Lai, R. W. Lau, and M.-H. Yang, "Learning fully convolutional networks for iterative non-blind deconvolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3817–3825, 2017.

[119] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.

[120] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *International Conference on Machine Learning*, pp. 5546–5557, 2019.

[121] C. Bertocchi, E. Chouzenoux, M.-C. Corbineau, J.-C. Pesquet, and M. Prato, "Deep unfolding of a proximal interior point method for image restoration," *Inverse Problems*, vol. 36, no. 3, p. 034005, 2020.

[122] C. Agarwal, S. Khobahi, A. Bose, M. Soltanalian, and D. Schonfeld, "DEEP-URL: A model-aware approach to blind deconvolution based on deep unfolded Richardson-Lucy network," in *IEEE International Conference on Image Processing*, pp. 3299–3303, 2020.

[123] S. Dai and Y. Wu, "Removing partial blur in a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2544–2551, 2009.

[124] J. Wu, X. Yu, D. Liu, M. Chandraker, and Z. Wang, "DAVID: Dual-attentional video deblurring," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 2376–2385, 2020.

[125] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.

[126] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *IEEE International Conference on Computer Vision*, pp. 3155–3164, 2019.

[127] H. Yue, C. Cao, L. Liao, R. Chu, and J. Yang, "Supervised raw video denoising with a benchmark dataset on dynamic scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2301–2310, 2020.

[128] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, "A convex approach for image restoration with exact Poisson–Gaussian likelihood," *SIAM Journal on Imaging Sciences*, vol. 8, no. 4, pp. 2662–2682, 2015.

[129] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11531–11539, 2020.

[130] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Understanding deformable alignment in video super-resolution," *arXiv preprint arXiv:2009.07265*, vol. 2, no. 3, 2020.

[131] S. Woo, J. Park, J.-Y. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision*, pp. 3–19, 2018.

[132] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 291–301, 2018.

[133] Y. Qiu, R. Wang, D. Tao, and J. Cheng, "Embedded block residual network: A recursive restoration model for single-image super-resolution," in *IEEE International Conference on Computer Vision*, pp. 4180–4189, 2019.

[134] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481, 2018.

[135] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[136] K. Purohit and A. Rajagopalan, "Region-adaptive dense network for efficient motion deblurring," in *AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11882–11889, 2020.

[137] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

[138] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, pp. 694–711, Springer, 2016.

[139] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[140] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[141] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*, pp. 2366–2369, IEEE, 2010.

[142] E. Katsaros, P. K. Ostrowski, D. Wesierski, and A. Jezierska, "Concurrent video denoising and deblurring for dynamic scenes," *IEEE Access*, vol. 9, pp. 157437–157446, 2021.

[143] T. Zhang, S. Song, Z. Jia, J. Yang, and N. K. Kasabov, "Object motion deblurring in single image under static background," *IEEE Access*, vol. 8, pp. 218069–218080, 2020.

[144] A. Levin, "Blind motion deblurring using image statistics," *Advances in Neural Information Processing Systems*, vol. 19, pp. 841–848, 2006.

[145] A. Chakrabarti, T. Zickler, and W. T. Freeman, "Analyzing spatially-varying blur," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2512–2519, IEEE, 2010.

[146] T.-L. Wang, K.-Y. Lee, and Y.-C. F. Wang, "Partial image blur detection and segmentation from a single snapshot," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1907–1911, IEEE, 2017.

[147] K. Purohit, A. B. Shah, and A. Rajagopalan, "Learning based single image blur detection and segmentation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2202–2206, IEEE, 2018.

[148] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren, "Spatio-temporal filter adaptive network for video deblurring," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2482–2491, 2019.

[149] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[150] E. Katsaros, A. Jezierska, and D. Wesierski, "Leveraging spatio-temporal features for joint deblurring and segmentation of instruments in dental video microscopy," in *2021 14th International Conference on Human System Interaction (HSI)*, pp. 1–5, IEEE, 2021.

[151] M. Zhang, Q. Gao, J. Wang, H. Turbell, D. Zhao, J. Yu, and Y. Lu, "RT-VENet: A convolutional network for real-time video enhancement," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4088–4097, 2020.

[152] G. Zhu, Z. Piao, and S. C. Kim, "Tooth detection and segmentation with mask r-cnn," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 070–072, IEEE, 2020.

[153] J. Kühnisch, O. Meyer, M. Hesenius, R. Hickel, and V. Gruhn, "Caries detection on intraoral images using artificial intelligence," *Journal of dental research*, 2021.

[154] U. Rashid, A. Javid, A. R. Khan, L. Liu, A. Ahmed, O. Khalid, K. Saleem, S. Meraj, U. Iqbal, and R. Nawaz, "A hybrid mask rcnn-based tool to localize dental cavities from real-time mixed photographic images," *PeerJ Computer Science*, 2022.

[155] J. F. Low, T. N. M. Dom, and S. A. Baharin, "Magnification in endodontics: A review of its application and acceptance among dental practitioners," *European journal of dentistry*, vol. 12, no. 04, pp. 610–616, 2018.

[156] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3614–3633, 2021.

[157] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, "Deep residual fourier transformation for single image deblurring," *arXiv preprint arXiv:2111.11745*, 2021.

[158] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proceedings of 1st International Conference on Image Processing*, vol. 2, pp. 168–172, IEEE, 1994.

[159] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[160] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "Simpleelastix: A user-friendly, multi-lingual library for medical image registration," in *IEEE CVPR Workshops*, pp. 134–142, 2016.

[161] T. Ehret, A. Davy, J.-M. Morel, G. Facciolo, and P. Arias, "Model-blind video denoising via frame-to-frame training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11369–11378, 2019.

[162] V. Nekrasov, T. Dharmasiri, A. Spek, T. Drummond, C. Shen, and I. Reid, "Real-time joint semantic segmentation and depth estimation using asymmetric annotations," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7101–7107, IEEE, 2019.

[163] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, 2019.

[164] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang, "Learning blind video temporal consistency," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 170–185, 2018.

[165] E. Katsaros, P. K. Ostrowski, K. Włódarczak, E. Lewandowska, J. Ruminski, D. Siupka-Mróz, Ł. Lassmann, A. Jezierska, and D. Węsierski, "Multi-task video enhancement for dental interventions," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*, pp. 177–187, Springer, 2022.

[166] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766, 2015.

[167] M. K. Hasan, L. Calvet, N. Rabbani, and A. Bartoli, "Detection, segmentation, and 3d pose estimation of surgical tools using convolutional neural networks and algebraic geometry," *Medical Image Analysis*, vol. 70, p. 101994, 2021.

[168] S. Lee, D. Cho, J. Kim, and T. H. Kim, "Restore from restored: Video restoration with pseudo clean video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3537–3546, 2021.